AN INVESTIGATION OF PEDAGOGICAL INTERVENTIONS WITHIN THE PRODUCTIVE FAILURE METHODOLOGY

A Dissertation
Presented to
The Academic Faculty

by

Dar-Wei Chen

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Psychology
in the School of Psychology, College of Sciences

Georgia Institute of Technology
May 2018

An Investigation of Pedagogical Interventions
Within the Productive Failure Methodology

Approved by:

Dr. Richard Catrambone, Advisor
School of Psychology
Georgia Institute of Technology

Dr. Mark Guzdial
College of Computing
Georgia Institute of Technology

Dr. Phillip Ackerman
School of Psychology
Georgia Institute of Technology

Dr. Rick Thomas
School of Psychology
Georgia Institute of Technology

Dr. Jamie Gorman
School of Psychology
Georgia Institute of Technology

Date Approved: March 9, 2018

TABLE OF CONTENTS

LIST OF TABLES

SUMMARY

The assistance dilemma asks how learning environments should "balance information or assistance giving and withholding" (Koedinger & Aleven, 2007, p. 239). Minimal guidance (MG) methods posit that students learn best when exploring problems freely, while direct instruction (DI) methods provide canonical solutions early on to streamline students' efforts (problems later). Each method type provides unique benefits, but both are important (Schwartz & Martin, 2004) and not easily delivered together. A relatively new MG-based method called "productive failure" (PF) is hypothesized to capture both sets of benefits by requiring students to struggle through problems early on and only revealing canonical solutions afterward (Kapur, 2008). Students using PF are hypothesized to more effectively transfer and retain information because balancing heuristics and formal knowledge produces diverse solution attempts (diSessa & Sherin, 2000) and struggling during exploration pushes students to identify and fill knowledge gaps (Kulhavy & Stock, 1989). In the present studies, participants learned to perform tasks in two domains, cryptarithmetic (more traditional) and Rubik's Cube (psychomotor, less traditional) while using either PF or DI methods. General linear models revealed that A) PF participants did not outperform DI participants on either immediate post-tests or retention tests, although they did report being more exploration-oriented during problem-solving and trying more unique solution strategies, B) subgoal labels increased learning, but only for the relatively novel Rubik's Cube domain (and they sometimes increased workload in the cryptarithmetic domain, in fact), C) the effects of subgoal labels did not change with instruction type, D) "testing effect" did not change across instruction type, but did change across domain. Future research is needed to determine how PF methods can be modified and/or scaffolded so that exploration mindsets and diverse solutions attempts help learners transfer and retain knowledge.

CHAPTER 1. INTRODUCTION

For many years, education researchers have debated a seemingly simple question called "the assistance dilemma," which can be summarized as: "How should learning environments balance information or assistance giving and withholding to achieve optimal student learning?" (Koedinger & Aleven, 2007, p. 239). The answer to this question has the potential to shape future instructional design in fundamental ways, but no consensus has been reached thus far. For now, two categories of instructional methods dominate the debate. Traditional methods that provide canonical instruction early on and utilize problem-solving as application practice are called "direct instruction" (DI), while "minimal guidance" (MG) methods require learners to discover information through guided exploration and problem-solving, instead of receiving canonical instruction.

Although MG and DI methods are pedagogically different, they are similar in that they both strive to help students avoid struggle and failure (i.e., being unsuccessful in producing canonical solutions) while learning; both types of methods provide various levels of scaffolding to reduce learner struggle and failure, ostensibly because struggle and failure ultimately do more harm than good. However, a relatively new method called "productive failure" (PF; e.g., Kapur, 2008) is hypothesized to leverage struggle and failure for unique learning benefits. In PF, learners attempt problems first before receiving canonical instruction and it is hypothesized that as a result, they will potentially be abler to A) solve transfer problems, B) retain knowledge past immediate comprehension tests, C) know *why* a given solution is correct, as opposed to just knowing *that* it is correct, and D) identify their own gaps in knowledge, among other benefits. Furthermore, given that PF is an exploration-based method with canonical instruction

implemented, learners using PF are hypothesized to reap benefits usually associated with minimal guidance (e.g., self-generated concepts) and direct instruction (e.g., streamlining of attention and resource allocation). Individual differences in learners will also play a role in how effective PF methods can be relative to DI, and some of these effects, related to pre-existing abilities and method structure levels in particular, will be discussed in the conclusions section of this document. The experiments described here tested the "productive failure" hypothesis and aim to provide new perspective to existing learner assistance approaches as well.

1.1 "Minimal guidance" model

Productive failure methods are based, in part, on a variety of existing minimal guidance methods, but PF is hypothesized to improve on each of those methods in some fashion.

1.1.1 Discovery learning

An early instantiation of minimal guidance was "discovery learning," in which students freely explore domains and material for themselves to create governing insights about the world (Anthony, 1973), often without concrete goals in mind. According to Bruner (1961), students learning this way are more likely to become "autonomous and self-propelled," as opposed to motivated solely by extrinsic factors such as grades. Furthermore, Bruner (1961) adds that teaching with an aim toward long-term learning is effective because "when behavior is long-range and competence-oriented, it comes under the control of more complex cognitive structures," freeing the behavior from "immediate stimulus control" (p. 6). The hypothesized effects of methods like discovery learning are more durable retention and better transfer to novel problems, sometimes at the expense of short-term boosts in performance (e.g., Dean & Kuhn, 2008). One disadvantage of using discovery methods is that due to the inherently low structure of the methods, the accuracy of information gained through exploration cannot be guaranteed; in

productive failure methods, canonical instruction is used to remedy that issue, albeit later in the learning process.

1.1.2 Constructivism

Learners in constructivist environments are hypothesized to build "conceptually functional representations of the external world" that are not necessarily unique to themselves (Jonassen, 1991, p. 61). Therefore, while the basic pedagogical premise of constructivism is similar to that of discovery learning (i.e., active construction of meaning), a conceptual difference is that in discovery learning, students are hypothesized to instead construct their own unique representations of the world. Per the cognitive theory of constructivism, common knowledge exists and students should be allowed to explore that common knowledge. For that reason, PF methods eventually use canonical instruction at some point, even if only after learners construct representations.

However, as an educational philosophy, the implementation of constructivism is often more closely-related to discovery learning in that learners are thought to create unique meanings for concepts (Guzdial, 1997). Jonassen (1991) observes that real-world contexts are perhaps best for learning via constructivism (the educational philosophy) and individualized meaning because connections made by learners will have a higher likelihood of being externally relevant, not limited and sidetracked by the bounds of a school environment. This observation is particularly important when students are creating unique meaning for learned information, Conceptually, these ideas are in line with theories such as situated cognition (e.g., Brown, Collins, & Duguid, 1988), which hypothesizes that learned information tends to be associated with the context in which it was learned and that appropriate contexts are therefore central to effective learning.

1.1.3 Impasse-driven learning

Impasse-driven learning is one of the first methods to implement struggle and failure to a large extent; impasses are defined by VanLehn et al. (2003, p. 220) as situations in which a student is stuck, "detects an error, or does an action correctly but expresses uncertainty about it." In contrast, discovery learning and constructivism, while actively engaging, are not as explicit about the use of struggle and failure. The governing principle of impasse-driven learning is that impasses are effective in helping learners adopt learning-oriented mindsets, which cause them to be more likely to search their memories, examine the environment, or ask nearby people, etc. in attempts to discover what they do not understand (VanLehn et al., 2003). The breakdown-driven learning method proposed by Winograd & Flores (1987) is conceptually similar.

After students reach impasses, tutors are to provide explanations soon after when students are not able to. This philosophy runs counter to that of productive failure, in which learners are encouraged to struggle perhaps a bit more, with instruction being delayed further and taking a canonical form (as opposed to non-canonical "just in time" explanations). Cope and Simmons (1994) write that providing feedback too soon can inadvertently shield learners from having to create high-level problem-solving strategies that they otherwise would be more likely to do if left to struggle some. Although learners can benefit some from impasse-driven learning, it is hypothesized that productive failure enables learners to achieve the full benefits of their struggles.

1.1.4 Effects on long-term knowledge and expertise

No matter the specific instantiation, MG methods are hypothesized to mitigate working memory constraints by encouraging learners to connect new information with prior long-term knowledge (Kapur & Bielaczyc, 2011) during the unstructured problem-solving periods. These

connections increase the chances that new information is understood at a deeper level than if it was learned via DI, where the new information is often stored in working memory and available in external memory. They also increase the storage strength (likelihood of later recall) of the new information, which is arguably more important in educational contexts than retrieval strength, which is merely a function of current activation and context cues (Bjork & Bjork, 1992).

Minimally-guided methods are also inherently more likely to help students learn how to structure problems independently, which is a trait indicative of expertise (Chi, Feltovich, & Glaser, 1981). In contrast, direct instructions often provide too much canonical information, especially early on, that students can then use as crutches to avoid having to structure problem spaces on their own.

1.2 "Direct instruction" model

Opposite minimal guidance in the learner assistance debate are direct instruction methods, which generally guide students strongly and limit exploration.

1.2.1 The "worked-example" effect and working memory

The worked example is considered "the epitome of strongly guided instruction" and "provides some of the strongest evidence for the superiority of directly guided instruction over minimal guidance" (Kirschner, Sweller, & Clark, 2006), p. 80). Worked examples are hypothesized to streamline attention to the most important parts of problems, reducing problem-solving search and thus lower working memory loads (Kirschner, Sweller & Clark, 2006). For most learners, and novices in particular, this streamlining is key because they do not possess the relevant schemas with which to integrate new information and prior knowledge, and therefore cannot construct new schemas that are durable (Rourke & Sweller, 2009). When unguided, many novices often resort to methods such as trial-and-error which are burdensome on working

memory, causing it to be unavailable for contributing to long-term memory (Kirschner, Sweller, & Clark, 2006). If working memory is occupied with tasks such as trial-and-error or problem-solving search, unguided students will not be able to use working memory to learn, and they could therefore potentially search problem spaces for long periods without adding to long-term memory (Sweller, Mawer, & Howe, 1982). Learners can also sometimes lean too much on pre-existing knowledge to explore a domain (as opposed to devising learning goals), which can then lead to flawed conclusions (Wineburg & Fournier, 1994). The positive effects of worked examples have been demonstrated by Sweller and Cooper (1985), for example, who found that students learned algebra more effectively when studying worked examples than when completing MG-style problem-solving.

Interestingly, hypotheses for both minimal guidance and direct instruction include lower working memory loads. In MG, learners are hypothesized to have lower working memory loads through reliance on long-term knowledge (e.g., Kapur & Bielaczyc, 2011); in DI, the reduction in problem-solving search is hypothesized to be better in reducing these loads (e.g., Kirschner, Sweller, & Clark, 2006).

1.2.2 Other evidence and arguments in favor of direct instruction

Direct instruction can be instantiated in many ways: Lectures, models, videos, presentations, demonstrations, as well as the aforementioned worked examples (Clark, Kirschner, & Sweller, 2012). They are all hypothesized to reduce misconceptions that can occur when learners receive minimal guidance (Brown & Campione, 1994) and therefore shield learners from mentally-taxing "false starts" (Carlson, Lundy, & Schneider, 1992). That is, encoding errors are less likely in DI environments and correct domain knowledge is more likely (Sweller & Chandler, 1991).

As opposed to MG methods in which problems are the mechanism for delivering content, problems in DI are used as opportunities for learners to practice applying canonical instruction, with scaffolding often "fading" over time in hopes that the learners can gradually become self-sufficient. This arrangement is hypothesized to be more optimal because when minimally-guided, students can sometimes have challenges distinguishing generalizable content knowledge from the specific contexts of the problem(s) they were given (Patel, Groen, & Norman, 1993). Therefore, transfer of learning between contexts might be difficult for some learners in MG environments. This issue is hypothetically remedied in DI by preventing learners from encountering problems until they have learned about the most relevant features of the given problem (Kirschner, Sweller, & Clark, 2006).

For all of these reasons, according to Mayer (2004), direct instruction has proven across many analyses to produce learning that is superior to minimal guidance. Furthermore, even on secondary measures such as engagement and frustration, some evidence exists that favors DI methods over "problem-solving prior" methods (e.g., Hardiman, Pollatsek, & Weil, 1986). However, in terms of productive failure specifically, not as many analyses have included the method because of its relative newness, so it remains to be seen whether these patterns in the data hold for PF.

1.3 Solving the assistance dilemma through "productive failure"

A growing body of literature posits that productive failure can improve student learning beyond what is usually achieved through minimal guidance or direct instruction (e.g., Kapur, 2011). This literature reveals that learners' struggles and failures are important on a cognitive level and can be leveraged to achieve learning objectives usually associated with MG or DI (i.e., the "MG vs. DI" debate might be a false choice).

A summary of MG and DI will be instructive in drawing comparisons with PF:

- *Minimal guidance* is characterized by the delivery of content mainly via problems, often with scaffolding to keep students from veering too far off-track. Learners are not explicitly presented with canonical instruction during a designated learning period for that purpose, although some or all of the canonical instruction content will be presented in the scaffolding.
- *Direct instruction* is characterized by the delivery of content through canonical instruction; problems are usually implemented for students to practice applying their new knowledge after using the canonical instruction.

On a high level, productive failure requires students to invent solutions to presented problems first (in the "generation period") before receiving canonical instruction ("consolidation period"), thereby reversing the traditional order of these two teaching elements in DI. This order leads to struggle (and ultimately, failure) early on in the learning process, but there often exists "a latent productivity in what initially seemed to be failure" (Kapur, 2008, p. 379). The generation effect, "which refers to the long-term benefit of generating an answer, solution, or procedure versus being presented that answer, solution, or procedure" (Bjork & Bjork, 2011), could explain this latent productivity, in part. The ensuing canonical instruction then serves to combat the "negative transfer" (Bransford & Schwartz, 1999) that often plagues minimal-guidance methods. It should be noted, however, that PF students do receive some basic domain information before entering the generation period, which lessens the probability of unproductive failures in which students attempt solutions that are too irrelevant to yield any valuable information.

Most MG methods employ scaffolding so that learners can avoid failure, ostensibly because it will hinder learning; however, failure is embraced and explicitly designed into the PF process through the use of problem-solving early on (generation period), and difficult ill-structured problems in particular are frequently used. In practice, scaffolding is withheld and "solution features" are deliberately made inconspicuous in PF so learners will be unlikely to guess canonical solutions, instead being encouraged to lean on heuristics and prior knowledge to generate solutions (Loibl & Rummel, 2014a). The focused "foraging for solutions" that occurs in PF can also be contrasted with discovery learning, in which learners usually explore without concrete objectives provided for them.

After initial problem-solving, canonical instruction follows for learners to fill in the rest of their understanding and remedy any mistakes they made. Sometimes, an initial assessment is implemented first immediately after the initial problems to ensure more concrete failure. Both exploration and canonical instruction are important (Schwartz & Martin, 2004), so this combination of the two is hypothesized to help learners achieve beyond what MG or DI alone can provide. Table 1 outlines the basic structures of MG, DI, and PF methods.

Table 1.

Basic structures of MG, DI, and PF methods

<table>
<tr><th></th><th>Phase 1</th><th>Phase 2</th><th>Phase 3</th></tr>
<tr><td>Minimal guidance</td><td colspan="3">Problem-solving (with scaffolding)</td></tr>
<tr><td>Direct instruction</td><td colspan="2">Canonical instruction</td><td>Problem-solving (with scaffolding)</td></tr>
<tr><td>Productive failure</td><td colspan="2">Problem-solving (no scaffolding), plus optional initial assessment</td><td>Canonical instruction</td></tr>
</table>

To concretely illustrate the difference between PF and DI, below are descriptions of both instructional methods with respect to a common average-speed problem (Kapur & Bielaczyc, 2012), holding total time-on-task constant across conditions.

- *PF conditions:* The story presented in the problem used dialogue to indirectly depict two people who were to reach a destination simultaneously under various constraints (different modes of transportation, waiting times, etc.). In the first period, students invented solution methods to this ill-structured problem; after an initial assessment consisting of similar problems, students went on to the consolidation phase in which they compared their invented solutions with canonical solutions and learned general concepts from canonical instruction.
- *DI conditions:* Canonical instruction and worked examples were received first, after which students worked on well-structured problems (i.e., "solution features" clearly presented) that were similar to the presented examples. An example of one of these problems is "Jack walks at an average speed of 4 km/hr for one hour. He then cycles 6 km at 12 km/hr. Find his average speed for the whole journey" (Kapur & Bielaczyc, 2012, p. 83).

Each of the following sections summarizes a key component of the productive failure hypothesis as it relates to math and physics (the domains that have been studied most so far with respect to PF).

1.3.1 Heuristics plus formal knowledge

In minimal-guidance environments, learners are led to utilize prior knowledge and heuristics during problem-solving, thereby mitigating some working memory constraints (Kapur & Bielaczyc, 2011) on the whole, even if searching problem spaces also increases learners'

working memory burdens somewhat (Sweller, 1988). In the event that some learners do encounter higher cognitive demands in PF, they also often report feeling more engaged because of the autonomy they are afforded during initial problem-solving (diSessa, Hammer, Sherin, & Kolpakowski, 1991). This prior knowledge activation is crucial for helping learners connect new material with long-term knowledge, which enables better encoding and assembling of schemas (Hiebert & Grouws, 2007) as well as better transferability and durability of learning (Kapur, 2012).

The blending of heuristics, prior knowledge, and formalized canonical instruction allows PF methods to provide benefits that MG and DI alone cannot. For example, PF students are more likely to generate relatively large amounts of diverse solutions for novel problems (diSessa & Sherin, 2000), a hallmark of how experts attempt problems (Clement, 1991; Reif & Larkin, 1991). In the aforementioned example problem regarding average speed, invented solutions from students include algebraic representations of the story, "brute-force" methods (guessing a distance and adjusting), diagrams, and conceptual statements about the variables (Kapur & Bielaczyc, 2012). Through these diverse solution attempts, students are expected to develop the ability to extrapolate new information to other contexts (procedural flexibility; Gorman, Cooke, & Amazeen, 2010). Another hypothesized benefit is the priming of students to solve transfer problems later using the relative wealth of available information (prior knowledge, heuristics, canonical instruction), even if the information is not germane to any given initial problem (Bransford & Schwartz, 1999).

A fair question regarding the above information might be whether DI methods can also achieve results similar to PF, given that many of them also implement canonical instruction and problem-solving. The key difference is that in productive failure, students use problem-solving to

"assemble or structure key ideas and concepts while attempting to represent and solve the ill-structured problems" (Kapur, Dickson, & Yhing, 2010, p. 1722). However, in direct instruction, problems are used "not as vehicles for making discoveries, but as a means of *practicing* recently-learned content and skills" (Clark, Kirschner, & Sweller, 2012, p. 6). As a result, students in DI are less likely to blend heuristics and formal knowledge, and more likely to receive formal knowledge and merely re-activate it when solving problems, leading to transferability that is not as robust. In contrast, PF students are led to use heuristics and prior knowledge during initial problem-solving (before receiving canonical instruction to remedy gaps in understanding), which ensures that both knowledge types are activated while learning. It is this "problem-solving prior" instructional order that enables students to adopt expertise- and mastery-oriented learning goals, which tend to produce durable learning (Belenky & Nokes-Malach, 2012) because of the deep structural knowledge necessary for expertise and mastery. Inventing solutions first, before receiving canonical instruction, also illuminates students' gaps in understanding much more readily than if they receive canonical instruction up front (DeCaro & Rittle-Johnson, 2012). The generation process, therefore, helps students to tailor their usage of the canonical solutions when they do arrive.

1.3.2 Failure-related cognition

"Expectation failure" is the idea that learning is most successful when the outcome expected by a student from the domain does not, in fact, occur (Schank, 1997). Key principles of expectation failure include:

- Learners are less likely to develop creative solution attempts if environment is too controlled and failures are therefore not possible

- Leaners are predisposed to explaining occurrences in the domain and adjusting their mental models to avoid being surprised by similar events
- For expectation failures to be most effective, they must occur during initial/practice problem-solving (more likely to be activated in future problems)

The key function of expectation failures is exposing learners to gaps in their understanding and eliciting learners' natural misunderstanding-induced curiosity in the material. In these situations, learners are more driven to fill knowledge gaps on their own (e.g., studying feedback), particularly when discrepancies between solution attempts and canonical solutions are wide (Kulhavy & Stock, 1989). Due to the "problem-solving prior" instructional order, PF methods are particularly conducive to learners producing initial solution attempts that are discrepant from canonical solutions. Expectation failures also disrupt learners' stability bias, the overconfident belief that currently-accessible information will remain just as accessible in the future (Kornell & Bjork, 2009).

Chi's (2000) theory of the imperfect mental model also accords with the notion that failure can be effective and essential for learning; in short, the theory states that learning is done through updates to one's own mental models and that self-explaining, in particular, is an efficient way for learners to update their own models according to their own needs. Of course, learners must recognize flaws in understanding first before updating their models, so PF methods often implement initial assessments after initial problem-solving to provide this opportunity (these initial assessments also allow for more concrete failures). Recognition of flawed understanding naturally directs learners to allocate mental resources to the content most relevant to addressing said flaws (Durkin & Rittle-Johnson, 2012). In direct instruction, learners merely apply

canonical instruction to problems, which is less likely to create realizations of potentially-flawed understanding.

Furthermore, when experiencing failures and ensuing canonical instruction, learners will also tend to identify reasons that a solution is plausible and why non-canonical solutions do not always work, which improves their capacity for transfer to novel situations (Kapur & Lee, 2009). Comparing invented solutions and canonical solutions aids in the encoding of critical conceptual features and selecting relevant problem-solving procedures, even when performing transfer tasks (Siegler, 2002). For example, when students were allowed to observe the consequences of entering incorrect spreadsheet formulas, as opposed to being corrected immediately upon entering an incorrect formula, they achieved higher scores on transfer tasks than immediately-corrected students (Mathan & Koedinger, 2003). Students in PF conditions will have opportunities to compare invented/failed solutions and canonical solutions, which are hypothesized to provide benefits beyond what the "regurgitative" processes in DI can provide. In general, PF is well-suited for inducing the failures that are vital for expertise and deeper learning. However, one caveat is that PF is likely to be most effective for domains in which fundamental and generalizable rules exist (e.g., STEM subjects) because failures in those domains can reveal essential structures that govern the domain and provide information beyond that particular instance. PF methods are less likely to be effective in domains in which rules are less generalizable and less connected. For example, one could imagine a video game in which the player uses a character to navigate a world (e.g., Super Mario) that possesses no set rules – a wall could be solid or illusory and in any given instance, it is impossible to predict based on prior instances. Exploration of such a world would reveal no enduring principles that the player could use to inform future actions in the video game.

1.3.3 Immediate performance vs. enduring learning

"Desirable difficulties" (Bjork, 2013), even if not severe enough to consistently induce failure, can still induce decreased immediate performance and PF-related learning benefits in the long term. Examples of these difficulties include:

- *Environmental factors*: When training to screen luggage, interface clutter can improve learning relative to clutter-absent training (Fiore, Scielzo, Jentsch, & Howard, 2006).
- *Training variation*: Practicing beanbag throws to targets of varying distances produces better performance than practicing at only the tested distance, even if the varying distances do not include the tested distance (Kerr & Booth, 1978)
- *Scheduling*: If a task comprises multiple components, interleaved practice scheduling (random practice order of components) produces better retention performance than blocked practice scheduling (practicing one component repeatedly until switching), even though improvement during training is slower (e.g., Shea & Morgan, 1979),
- *Secondary tasks*: When training on a radar detection primary task, adding an irrelevant concurrent secondary task decreased primary task test performance, even when the test itself included the irrelevant secondary task. However, test performance increased with the addition of a third task during training (a relevant concurrent secondary task, in addition to the primary and irrelevant secondary tasks), even when just the primary task and irrelevant secondary task were included on the test (Young et al., 2011).

Young et al. (2011) additionally emphasize that only task-relevant difficulties can be "desirable" and produce improved learning.

The goal of any instructional method should be learning, which can be defined as "permanent changes in comprehension, understanding, and skills of the types that will support long-term retention and transfer" (Soderstrom & Bjork, 2015, p. 176). Learning is a separate observed variable from immediate performance, which is a possibly temporary measure that can be an unreliable indicator of learning (Soderstrom & Bjork, 2015). Many instructional methods focus on producing immediate performance improvements, but some evidence indicates that immediate performance is not indicative of long-term retention and/or transfer, which is perhaps more important (e.g., Schmidt & Bjork, 1992). For example, in PF, students invariably learn relatively slowly early on as they struggle while exploring the domain, but often surpass their DI counterparts later on even with study time held constant.

When learners demonstrate strong immediate performance, they could be merely exhibiting retrieval strength, which is recall activated in particular contexts; however, durable learning is a function of storage strength, which comprises the depths to which the material is associated with prior knowledge (Bjork & Bjork, 1992). Increasing storage strength is most efficiently done through information retrieval (as opposed to information review) because the creation of "new routes" to information inherently activates previous knowledge as well (Carrier & Pashler, 1992). It follows that productive failure, in which students learn in part by attacking problems in a variety of ways, might be more conducive to enduring learning than direct instruction. The fact that 84% of students report using re-reading as a key study strategy (Karpicke, Butler, & Roediger, 2009) is a testament to how ingrained repetition-driven DI

methods are in traditional instructional methods, but not necessarily an indication of the effectiveness of those methods.

The observation that enduring learning and immediate outward performance improvement can be uncorrelated is seen in research ranging from maze rats (rats' abilities to finish mazes improve after ostensibly random wandering; Blodgett, 1929) to statistics classes (students who invented solutions and received canonical instruction later outperformed DI students; Schwartz & Martin, 2004). Furthermore, methods that aim to improve immediate performance can actually undermine enduring learning: For example, frequent and/or specific feedback, a common DI component, often helps students complete test problems that are similar to the ones they practiced, especially if tested soon after instruction. However, learners that receive the crutch of immediate and frequent feedback are shielded from creating generalizable problem-solving strategies, an important skill that is developed in those that are forced to struggle without immediate feedback (Cope & Simmons, 1994). Kulik and Kulik (1988) found in their meta-analysis that, indeed, learners using delayed feedback performed 0.44 standard deviations better on retention tests than learners who received immediate feedback. Feedback that is too granular can also produce short-term performance increases that placate learners into less self-assessment than they would otherwise engage in (Goodman, Wood, & Hendrickx, 2004).

On a more general level, instructional methods that increase immediate and near-transfer performance should not be interpreted necessarily as methods that improve enduring and transferable learning. Productive failure is hypothesized to be a method that improves the latter at the expense of the former.

1.4 Examining subgoal scaffolding in productive failure

Many of the PF studies to this point have required learners to complete initial problem-solving (the "generation period") without scaffolding of any sort, perhaps because this arrangement increases the chances of failure and the learner reaping the benefits associated with failure. When no scaffolding structure is present, one potential concern is that learners might not fail in constructive ways, which could then lead to difficulty during canonical instruction because learners will have strayed "off course" to varying extents. Therefore, it is possible that PF methods could be even more optimal for learning with the implementation of some scaffolding, especially those scaffolding mechanisms that provide just enough guidance to ensure that failures are indeed productive (i.e., help students unearth fundamental truths about the domain). After all, according to Anthony (1973), scaffolding in some form is often necessary for minimally-guided methods.

A few PF studies have implemented scaffolding during the generation period, but there are many more scaffolding mechanisms to be examined with regards to interactions with PF, some of which might produce better learning than non-scaffolded PF methods. The scaffolding mechanism chosen for manipulation in the presently proposed study is "subgoals," which are labels for functional groupings of steps that can help learners recognize fundamental components of a problem (Catrambone, 1998). Subgoals are a promising scaffolding mechanism for PF because they can potentially alleviate one of the major weaknesses in PF methods, which is the possibility that learners might fail unproductively by misunderstanding the deep structure of a given problem space. That is, if a learner is not aware of fundamental objectives required to solve a problem, he or she will possibly perform actions that are irrelevant to learning the task, inducing frustration and perhaps unproductive failures (or "false starts"; Carlson, Lundy, &

Schneider, 1992). Subgoals can increase the likelihood of any given learner action being at least somewhat relevant to the problem, but because they do not directly instruct the detail-level mechanics of solving the problem, learners are still likely to fail in the generation of canonical solutions and therefore make the gains associated with such failures.

1.5 General overview of proposed study and hypotheses

The experiments in the present study compared the effectiveness of productive failure and direct instruction in two domains that have not been examined before in this PF context. In Experiment 1, participants learned about cryptarithmetic, a domain that functions like the traditional academic domain of algebra and is somewhat similar to physics and math domains that have been used in past PF studies, but is more likely to be unfamiliar to participants. The tasks inherent in this domain (deducing variable values, logical reasoning, etc.) allow for reasonable comparison of the results to those from existing PF studies, which have centered mostly on STEM domains. In Experiment 2 (which was procedurally identical to Experiment 1), participants learned about the Rubik's Cube, a spatially-oriented task that requires some psychomotor coordination. The generalizability of PF methods to non-traditional domains were tested in this experiment. Table 2 below lists some relevant differences between the two domains used in the current experiments:

Table 2.

Comparison of cryptarithmetic and Rubik's Cube on relevant dimensions

Cryptarithmetic	Dimension	Rubik's Cube
Minimal	Motor/spatial component	Substantial
Very similar	Similarity to traditional school subjects	Not similar
Unlikely, but possible	Can be solved through "brute force"	Almost impossible
Yes	Permanent external memory of work	No

It was hypothesized that in general, on measures of transfer and/or long-term retention, participants learning from productive failure would outperform those learning from direct instruction; the diversity of invented solutions during the PF "generation period," driven by the combination of heuristics and formal knowledge that is more likely with PF, were expected to enable learners to solve a larger variety of novel problems (diSessa & Sherin, 2000), and the activation of long-term knowledge more inherent during PF exploration was expected to support connections between new and old information that last beyond immediate post-tests (e.g., Bjork & Bjork, 1992). However, on near-transfer assessments administered immediately after learning, DI and PF were expected to produce similar performance because the aforementioned factors are less likely to matter when learners are required to merely reproduce the procedures they learned very recently.

As for the effects of subgoal labels, it was hypothesized that in accordance to the existing literature, they would improve participants' scores on all types of tasks because subgoals used while learning are almost always valuable for tasks in the same domain, even novel tasks that are not isomorphic to prior learned tasks. Catrambone (1998) suggests that when learners discover that the old steps within a subgoal do not work for a particular novel problem, learners benefit from knowing that they only have to consider changes to these particular steps in order to continue making progress toward the overall task goal (a "reduced search space"), as opposed to having to consider all of the steps in a task when it is not decomposed into subgoals. When subgoals are not provided, learners often default to memorizing steps while studying, which leads to performance decrements when attempting problems that are even slightly different from the studied problems (Reed et al., 1990); learners that merely memorize steps are also less likely to know what pre-existing knowledge could be helpful in any given task, an issue that is not as

prevalent with learners who recognize deeper subgoal structures (Catrambone, 1998). However, perhaps more interestingly, the magnitudes of these effects were also hypothesized to interact with the instructional manipulations: subgoals were expected to be much more helpful in PF conditions than in DI conditions because in PF, subgoals serve as the only organizing features to learners during their initial struggles while simultaneously not providing too much detail that would significantly decrease the chance of failures. Subgoals should increase the chances of learners failing productively, as opposed to "floundering" unproductively. In contrast, direct instruction materials already possess at least some level of structure that learners can use (e.g., order of steps), making any subgoal-imposed structures relatively less important, even if still useful in an absolute sense.

The experiments also incorporated a built-in examination of the "testing effect," which is "the finding that retrieval of information from memory produces better retention than restudying the same information for an equivalent amount of time" (Roediger & Butler, 2011, p. 20). Half of the participants received assessments both immediately after learning (post-test) and one week after learning (retention test), while the other half received just the retention test and extra study time equivalent to the time needed for the post-test. The testing effect was predicted to manifest itself as a main effect, with participants given the post-test performing better on the retention tests than those merely given more study time, likely due to the fact that retrieval practice improves how strongly any given retrieval cue is associated with a relevant piece of information (Karpicke & Blunt, 2011), or that the act of retrieval naturally leads to the creation of more routes to information and therefore increases the likelihood of any given route being activated when the information is needed (Carpenter, 2009). However, the presence of post-tests was also predicted to produce differential effects on retention depending on whether learners were using

PF or DI; more specifically, because PF learners were hypothesized to work relatively hard initially to retrieve information on immediate post-tests that DI learners could simply "regurgitate" (especially for procedurally-similar problems), they were also seen as inherently likelier to activate more related pieces of information that would create long-term pathways, thereby increasing the chances of a target concept being successfully retrieved in later retention tests (Carpenter, 2009).

A fourth manipulation in the current experiments was that of constraints on study time. Half of participants were assigned to conditions with extended study time, while the other half were assigned to conditions with limited study time. According to Kehoe, Stasko, and Taylor (2001), when learners are relatively unconstrained by time limits, they tend to engage in more exploratory activities. Therefore, given that PF is a methodology predicated on learner exploration (and the benefits that are associated with failure), it was hypothesized that scenarios with extended time would increase performance for PF learners more than for DI learners. Implementing a condition with extended study time also ensured that any relevant learning differences between any of the aforementioned manipulations have a higher likelihood of becoming apparent. After all, with limited study time, learners might not have sufficient opportunities to glean all of the available benefits from a given instructional method.

As mentioned previously, the second experiment in this study was constructed similarly to Experiment 1 except for the fact that the participants learned to solve part of the Rubik's Cube instead of cryptarithmetic problems. Experiment 2 provided an opportunity to examine whether Experiment 1 findings replicated or whether the effects of the manipulations might depend on how academic in nature the domain is. The specific methodological details that used to test these hypotheses, and the ones mentioned previously, can be found in the next chapter.

A summary of the tested hypotheses can be found below in Table 3:

Table 3.

Summary of hypotheses tested in the present studies

Number	Description
H1	Medium- and far-transfer problems, immediate post-test: PF participants will score more highly than DI participants (no difference for near-transfer problems)
H2	Retention test problems: PF participants will score more highly than DI participants
H3	Number of identified knowledge gaps, mid-point check: PF participants will identify more gaps than DI participants
H4	Mental workload (TLX), mid-point check: PF participants will report higher workload than DI participants
H5	Amount of prior knowledge and intuition used, mid-point check: PF participants will list more concepts than DI participants
H6	Near-transfer problem, mid-point check: PF participants will score lower than DI participants
H7	Role of problem-solving in learning process, post-learning: PF participants will be more likely to report that problem-solving was used for exploration while DI participants will be more likely to report that it was used for practice/application
H8	Identifying potential mistakes of future participants, post-learning: PF participants will identify more potential mistakes than DI participants
H9	Number of unique solution strategies invented, post-learning: PF participants will use more unique solution strategies than DI participants
H10	Perceived difficulty of material, post-learning: PF participants' subjective levels of difficulty reported will be higher than those reported by DI participants
H11	Mental workload (TLX), post-learning: PF participants will report higher workload than DI participants
H12	All problem types: Participants receiving subgoals will score more highly than participants without subgoals
H13	Mental workload (TLX), mid-point check and post-learning: Participants receiving subgoals will report lower subjective workload than participants without subgoals
H14	All problem types: Subgoals will improve performance for PF participants more than they improve performance for DI participants
H15	Testing effect: Presence of immediate post-test will increase retention scores of PF participants more than those of DI participants
H16	Time constraints: Performance improvements produced by extended time will be larger for PF participants than for DI participants

CHAPTER 2. METHOD – EXPERIMENT ONE (CRYPTARITHMETIC)

2.1 Participants

A meta-analysis of productive failure studies (Chen, 2016) found that PF methods have produced, on average, a performance improvement of about 0.66 SD in deep conceptual knowledge when compared to direct instruction methods, and because PF was hypothesized to improve this kind of generalizable knowledge (as opposed to performance on procedurally-similar tasks), this effect size drove the power analysis used to determine the sample size in this study. To achieve 80% power and 5% Type I error rate when searching for an effect of this size, 64 participants were used. These participants were recruited through the online SONA research participation system at the Georgia Institute of Technology and compensated with class credits for their time. All students at the Institute qualified for the experiment except for those who had prior experience in systematically solving cryptarithmetic problems.

2.2 Experimental design

Experiment 1 was a laboratory experiment in which all participants were required to learn how to solve basic cryptarithmetic addition problems involving two numbers. The four manipulated independent variables were:

- Instruction type (two levels, between subjects): productive failure or direct instruction)
- Subgoal labels (two levels, between subjects): subgoal labels were provided or withheld

- Immediate post-test (two levels, between subjects): participants were either given a post-test immediately after the study periods or a corresponding amount of additional study time
- Time limit (two levels, between subjects): participants were allocated either 10 minutes (limited time) or 20 minutes (extended time) during the two study periods

These four variables were fully crossed to form a factorial design for the experiment. Observed dependent measures included immediate task performance (near transfer, medium transfer, far transfer), retention task performance (near transfer, medium transfer, far transfer), and several secondary assessments that could predict task performance (e.g., number of solution methods generated, workload; see 2.3.6 for full list of secondary assessments and explanations). Learner characteristics (demographics and pre-existing ability in the domains) were also collected for examination as potential predictors of performance.

2.3 Materials and procedures

The procedures outlined next were fully executed, for each participant, in the presented order. Up to four people participated at once in the laboratory's computer workstation areas. Table 4 contains a high-level outline of Experiment 1 procedures in the cryptarithmetic domain, and details about each period and associated materials are summarized in the sections following the tables.

Table 4.

Cryptarithmetic experimental domain outline (Experiment 1)

Period	Direct instruction	Productive failure
0 (5min)	Demographics paperwork, consent form	Demographics paperwork, consent form
1 (3min)	Introduction to cryptarithmetic	Introduction to cryptarithmetic
2 (10/20min)	Canonical instruction (participant can take notes) [presence of subgoals depending on condition]	Problem-solving, solution generation (participant can take notes) [presence of subgoals depending on condition]
3 (13min)	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve cryptarithmetic addition problem	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve cryptarithmetic addition problem
4 (10/20min)	Problem-solving, solution generation (participant can use notes from Period 2) [presence of subgoals depending on condition]	Canonical instruction (participant can use notes from Period 2) [presence of subgoals depending on condition]
5 (6min)	Post-learning questions • What is purpose of Period 4? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance	Post-learning questions • What is purpose of Period 2? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance
6 (20min)	Immediate post-test [in "no post-test" conditions, provide more study time instead] • Addition, 2 numbers (near transfer) • Subtraction (medium transfer) • Multiplication (far transfer)	Immediate post-test [in "no post-test" conditions, provide more study time instead] • Addition, 2 numbers (near transfer) • Subtraction (medium transfer) • Multiplication (far transfer)
ONE-WEEK BREAK		
7 (20min)	Retention test • Addition, 2 numbers (near transfer) • Subtraction (medium transfer) • Multiplication (far transfer)	Retention test • Addition, 2 numbers (near transfer) • Subtraction (medium transfer) • Multiplication (far transfer)
8 (15min)	Logic/algebra ability test	Logic/algebra ability test

2.3.1 Period 0: Demographics paperwork (5 minutes)

When participants first arrived at the experimental facility, they were asked to complete a demographics questionnaire which included questions about gender, age, major, year in college, GPA, standardized test scores, familiarity with the domain, and the learner's personal expectation regarding the difficulty of learning in this domain. This questionnaire was administered on paper and the data was analyzed later for predictive power regarding task performance (this questionnaire can be seen in Appendix A).

2.3.2 Period 1: Introduction to domain (3 minutes)

Next, before the learning activities begin, participants received a short printed primer on the domain to provide context for the ensuing material. For the cryptarithmetic domain (see Appendix B), this primer included the overall learning objective (to solve an addition cryptarithm) and fundamental rules of the domain; participants were allocated three minutes to read this primer.

2.3.3 Period 2: First learning session (10 or 20 minutes, depending on condition)

This period was the first one in which participant procedures differed based on the assigned learning condition:

- Direct instruction: Participants in DI conditions received canonical instructions during this period, which comprehensively summarized the steps to solving the given primary task in the domain (cryptarithmetic: solving an addition problem). They were allowed to use the necessary materials to practice the given tasks, such as pencil and paper to solve cryptarithms. These canonical instructions were presented with or without subgoals depending on assigned condition. Appendix C provides an example of canonical instructions for cryptarithmetic with and

without subgoal labels; subgoals include: write the problem out, search for zeros and nines, search for 1s, generate and test.

- Productive failure: This period served as the "generation period" in which participants invented their own primary task solutions by problem-solving without the aid of canonical instructional materials. Participants assigned to non-subgoal conditions received no materials of any kind during this period, while those assigned to subgoal conditions received just high-level outlines of the steps to complete the primary tasks. The subgoal outline for completing cryptarithmetic addition problems are shown in Appendix D.

2.3.4 Period 3: Mid-point performance check (13 minutes)

After the first learning session, each PF and DI participant completed the same mid-point performance check, which served two key purposes: Ensuring that PF participants failed concretely in a test-like situation (to increase the chances of reaping the benefits of failure) and enabling the researchers to test the notion that early performance is not necessarily predictive of enduring learning. In the cryptarithmetic domain, participants were asked to solve an addition cryptarithm within five minutes (e.g., OOOH + FOOD = FIGHT)

However, in addition to completing a check on performance, participants also answered a few questions that examined the hypothesized benefits supposedly received by PF learners and not DI learners. These questions aimed to identify:

- The variety of solution strategies used by a given participant, including reasonable strategies that could turn out to be incorrect but demonstrate learning progress, but excluding strategies that very obviously did not demonstrate any learning

(e.g., “all instances of ‘N’ equal 4”). Each distinct strategy was recorded by two graders.

- How aware a given participant is of any knowledge gap he or she might have (number of distinct gaps identified were counted by two graders)
- Amount of prior knowledge and intuition/heuristics used by a given participant (each distinct idea was counted by two graders)
- A given participant’s prediction of his or her own performance level at this point
- The amount of workload thus far, as recorded on four NASA TLX scales

The full questionnaire is shown in Appendix E and participants completed it in paper form before attempting the aforementioned domain task.

2.3.5 Period 4: Second learning session (10 or 20 minutes, depending on condition)

The second learning session was similar to the first (described in section 2.3.3) except for the fact that the DI participants engaged in problem-solving (analogous to the “generation period” in productive failure; see Appendix D for the subgoal outline that was given to those in subgoal-related conditions) and PF participants received canonical instruction (see Appendix C for these instructions with and without subgoals implemented). This arrangement allowed participants of both instructional methods to receive the same material but in reverse order (DI: canonical instruction, then problem-solving; PF: problem-solving, then canonical instructions). In this second learning session, participants were also permitted to use any notes that they recorded from the first learning session.

2.3.6 Period 5: Secondary assessments (6 minutes)

When participants finished the second learning session, they answered some more secondary questions during this period before moving onto the primary assessments. Like the

questions from the mid-point performance check (period 3), these questions aimed to test the hypotheses of the productive failure instructional method. Participants were asked about:

- What they believed to be the pedagogical purpose of the problem-solving periods (period 2 in PF, period 4 in DI)
- Potential mistakes they would like to warn future participants about making (a "cover story" to elicit useful responses from the current participants), of which each distinct identified potential mistake was recorded in the data by two graders
- Perceived difficulty of the material on a 7-point Likert scale
- A given participant's prediction of his or her own performance level at this point
- The amount of workload thus far, as recorded on four NASA TLX scales

These questions were answered on a printed form that can be seen in Appendix F.

2.3.7 Period 6: Primary immediate learning assessments or extra study time (20 minutes)

At this point, all learning was complete and only learning assessments remained. However, to examine the "testing effect" on retention performance, just half of the participants were required to complete the learning assessments in this period while the other half were given an equivalent amount of time to further study all of the previously provided materials (see previous sections for these materials). If applicable to the condition, the assessments in this period (directions for post-test and the cryptarithms used for the post-test are shown in Appendix G) took place immediately after period 5 and covered:

- Near-transfer problem (similar to learned problem, five minutes): Solve an addition cryptarithm with two numbers
- Medium-transfer problem (small extrapolation needed from learned problem, six minutes): Solve a subtraction cryptarithm

- Far-transfer problem (large amount of extrapolation needed from learned problem, eight minutes): Solve a multiplication cryptarithm

2.3.8 Period 7: Primary retention learning assessment (20 minutes, depending on condition)

Although not every participant completed the immediate learning assessments described in the previous section, every participant completed retention learning assessments one week after the first experiment block (Period 0 through Period 6). The problems presented during this period were similar in nature to the ones in the immediate assessments, but tested the durability of participants' learning by virtue of the participants being required to wait one week before being tested on these problems. Retention performance is arguably a better indicator of learning than immediate performance and is also one of the areas in which PF methods are predicted to produce superior outcomes compared to DI methods. The problems presented on the retention test included (directions for retention test and the cryptarithms used for the retention test are shown in Appendix H):

- Near-transfer problem (similar to learned problem, five minutes): Solve an addition cryptarithm with two numbers
- Medium-transfer problem (small extrapolation needed, six minutes): Solve a subtraction cryptarithm
- Far-transfer problem (large amount of extrapolation needed, eight minutes): Solve a multiplication cryptarithm

2.3.9 Period 8: Test of relevant pre-existing abilities (15 minutes)

The last period of the experiment, held right after the retention test, was used to assess the pre-existing abilities of the learners that were relevant to the experimental domain. These tests were completed at the end of the experiment, as opposed to before, to A) avoid the possibility of

the tests "priming" participants before they start learning, and B) prevent the participants from inadvertently learning about the domains from the tests. Results from the tests were used to measure the effects of relevant pre-existing abilities on learning and performance. Summaries of the tests are below:

- Systems of equations (two 3-minute sessions with ten problems each): Participants solve problems consisting of systems of two equations (see Appendix I for example problems)
- Logic puzzles (one 6-minute session): Participants were provided with a story in which clues were given regarding the state of affairs and they were to answer questions about logical conclusions that could be drawn from this information (see Appendix I for an example question)

After these tests, participants were debriefed and provided the agreed-upon class credit for experiment participation.

2.4 Grading schemes for primary cryptarithmetic learning tasks

Every problem in the cryptarithmetic domain, regardless of problem type (near transfer, medium transfer, far transfer) or timing (immediate, retention) was graded in the same way. Participants received points based on the number of letters whose values they could decipher: A single correct answer for a letter was scored as 1 point, whereas a letter that was narrowed down to two value possibilities (including the correct value) was scored as a half-point. The scores for each problem were recorded as a percentage of the total points available for that problem (variable number of points available for each problem, depending on number of letters).

CHAPTER 3. METHOD – EXPERIMENT TWO (RUBIK'S CUBE)

3.1 Participants

All specifications of the participants in Experiment 2 were identical to those used in Experiment 1; 64 participants, who had no systematic experience with Rubik's Cube and cryptarithmetic, were recruited through Georgia Tech's SONA system and awarded class credits for their time.

3.2 Experimental design

All manipulations (instruction type, presence of subgoal labels, and presence of immediate post-test), dependent measures (immediate task performance, retention task performance, secondary assessments), and collected learner characteristics were the same as those used in Experiment 1.

3.3 Materials and procedures

Table 5 contains a high-level outline of Experiment 2 procedures in the Rubik's Cube domain, and details about each period and associated materials are summarized in the sections following the table.

Table 5.

Rubik's Cube domain experimental outline (Experiment 2)

Period	Direct instruction	Productive failure
0 (5min)	Demographics paperwork, consent form	Demographics paperwork, consent form
1 (3min)	Introduction to Rubik's Cube	Introduction to Rubik's Cube
2 (10/20min)	Canonical instruction (participant can take notes) [presence of subgoals depending on condition]	Problem-solving, solution generation (participant can take notes) [presence of subgoals depending on condition]
3 (9min)	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve first layer	Mid-point check • Knowledge gap identification • Engagement/curiosity/frustration/TLX • What prior knowledge/intuition did you use while learning (if any)? • What solution methods have you thought of during this period? • Prediction of performance • Solve first layer
4 (10/20min)	Problem-solving, solution generation (participant can use notes from Period 2) [presence of subgoals depending on condition]	Canonical instruction (participant can use notes from Period 2) [presence of subgoals depending on condition]
5 (6min)	Post-learning questions • What is purpose of Period 4? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance	Post-learning questions • What is purpose of Period 2? • What are potential mistakes that other participants after you might make? • Engagement/curiosity/frustration/TLX • How difficult is this material? • Prediction of performance
6 (15min)	Immediate post-test [in "no post-test" conditions, provide more study time instead] • First layer, yellow (near transfer) • First layer, green (medium transfer) • American flag mini (far transfer)	Immediate post-test [in "no post-test" conditions, provide more study time instead] • First layer, yellow (near transfer) • First layer, green (medium transfer) • American flag mini (far transfer)
ONE-WEEK BREAK		
7 (15min)	Retention test • First layer (near transfer) • First layer, red (medium transfer) • French flag mini (far transfer)	Retention test • First layer (near transfer) • First layer, red (medium transfer) • French flag mini (far transfer)
8 (15min)	Spatial ability test	Spatial ability test

3.3.1 Period 0: Demographics paperwork (5 minutes)

The questionnaire used in Experiment 2 was the same as the questionnaire used in Experiment 1 (see Appendix A).

3.3.2 Period 1: Introduction to domain (3 minutes)

Next, before the learning activities begin, participants received a short printed primer on that domain to provide context for the ensuing material. For the Rubik's Cube domain, this primer (see Appendix J) included information about the physical structure of the cube, the overall learning objective (to solve the first layer from the "yellow side"), and how to use the tutorial; participants were allocated three minutes to read this primer.

3.3.3 Period 2: First learning session (10 or 20 minutes, depending on condition)

This period was the first one in which participant procedures differed based on the assigned learning condition:

- Direct instruction: Participants in DI conditions received canonical instructions during this period, which comprehensively summarized the steps to solving the given primary task in the domain (Rubik's Cube: solving the "first layer" on the yellow side). They used computers to interact with the tutorials and were also given the necessary materials to practice the given tasks (i.e., an actual Rubik's Cube). These canonical instructions were presented with or without subgoals depending on assigned condition. Appendix K provides an example of canonical instructions for the Rubik's Cube (with and without subgoal labels; subgoals include: make a cross, rotate the cross, swap the incorrect cross pieces, insert the four bottom corners).

- Productive failure: This period was the "generation period" in which participants invented their own primary task solutions by problem-solving without the aid of canonical instructional materials. Participants assigned to non-subgoal conditions received no materials of any kind during this period, while those assigned to subgoal conditions received just high-level outlines of the steps to complete the primary tasks. The subgoal outline for completing the first layer on the Rubik's Cube is shown in Appendix L.

3.3.4 Period 3: Mid-point performance check (9 minutes)

After the first learning session, each PF and DI participant completed the same mid-point performance check; in the Rubik's Cube domain, participants were asked to solve the first layer (on the yellow side) within five minutes. However, in addition to completing a check on performance, participants also answered a few questions that examined the hypothesized benefits supposedly received by PF learners and not DI learners. These questions were the same as the ones described in section 2.3.4 and can be seen in Appendix E.

3.3.5 Period 4: Second learning session (10 or 20 minutes, depending on condition)

The second learning session was similar to the first (described in section 3.3.3) except for the fact that the DI participants engaged in problem-solving (analogous to the "generation period" in productive failure; see Appendix L for the subgoal outlines given to those in subgoal-related conditions) and PF participants received canonical instruction (see Appendix K for these instructions with and without subgoals implemented). This arrangement allowed participants of both instructional methods to receive the same material but in reverse order (DI: canonical instruction, then problem-solving; PF: problem-solving, then canonical instructions). In this

second learning session, participants were also permitted to use any notes that they recorded from the first learning session.

3.3.6 Period 5: Secondary assessments (6 minutes)

When participants finished the second learning session, they answered some more secondary questions during this period before moving onto the primary assessments. These questions were the same as the ones described in section 2.3.6 and can be seen in Appendix F.

3.3.7 Period 6: Primary immediate learning assessments (15 minutes)

Half of the participants completed an immediate learning assessment covering the following items (see Appendix M for post-test directions) while half of the participants were given an equivalent amount of study time to further review all of the materials provided previously:

- Near-transfer problem (similar to learned problem, four minutes): Solve the first layer on the yellow side from a newly-scrambled cube
- Medium-transfer problem (small extrapolation needed from learned problem, five minutes): Solve the first layer on the green side from a newly-scrambled cube (conceptually similar to first layer on yellow side, but requires re-thinking of color matches)
- Far-transfer problem (large amount of extrapolation needed from learned problem, six minutes): Create a miniature version of the American flag from a newly-scrambled cube, a design whose execution does not directly follow necessarily from principles learned during the tutorial.

3.3.8 Period 7: Primary retention learning assessment (15 minutes)

Although not every participant completed the immediate learning assessments described in the previous section, every participant completed retention learning assessments one week after the first experiment block (Period 0 through Period 6). The problems presented on the retention test included (directions for retention test and pictures of solved states are shown in Appendix N):

- Near-transfer problem (similar to learned problem, four minutes): Solve the first layer on the yellow side from a newly-scrambled cube
- Medium-transfer problem (small extrapolation needed, five minutes): Solve the first layer on the red side from a newly-scrambled cube (conceptually similar to first layer on yellow side, but requires re-thinking of color matches)
- Far-transfer problem (large amount of extrapolation needed, six minutes): Create a miniature version of the French flag from a newly-scrambled cube, a design whose execution does not directly follow necessarily from principles learned during the tutorial.

3.3.9 Period 8: Test of relevant pre-existing abilities (15 minutes)

The last period of the experiment, held right after the retention test, was used to assess the pre-existing abilities of the learners that are relevant to the experimental domains. Summaries of the tests are below and the associated appendices are referenced:

- Paper-folding (two 3-minute sessions with ten problems each): A paper is folded and a hole is punched – the participant is to figure out where the holes are in the paper when it is completely unfolded (see Appendix O for example problem)

- Cube-folding comparison (one 6-minute session with ten problems): A cube is formed by folding six faces together – the participant was to figure out which presented cube arrangement is impossible given the faces folded together (see Appendix O for example problem)

After these tests, participants were debriefed and provided the agreed-upon class credit for experiment participation.

3.4 Grading schemes for Rubik's Cube primary learning tasks

Points were awarded on post-tests and retention tests based on participants' abilities to place relevant pieces in the correct locations, with an emphasis on edge pieces (the pieces that compose "the cross") because of the difficulty of placing those pieces and the fundamental nature of those pieces. The guiding principles of these schemes have been used in previous experiments (Chen & Catrambone, 2016; Chen & Catrambone, 2014) and were based in part on correspondence with 2007 Florida Open Rubik's Cube champion Andrew Chow (A. Chow, personal communication, May 11, 2015).

- Near-transfer problems (immediate and retention): A three-tiered scheme, in accordance with the three subgoals of solving the Rubik's Cube first layer, was implemented in which participants could receive points in the next tier only if the previous tier was completed; it was structured this way because pieces in the next tier could potentially be completed accidentally while solving the previous tier. Tier 1 covered the creation of the cross without regards to matching centers on the adjacent sides (third and fourth edge pieces weighted more heavily because placing them is more difficult when the first two edge pieces are already inserted),

Tier 2 covered the matching of those centers, and Tier 3 covered the correct insertion of corner pieces. Table 6 below outlines this scoring scheme:

Table 6.

Rubik's Cube scoring scheme: Near- and medium-transfer problems

	Cross pieces in place	Matching centers	Corner pieces in place	Score
Tier 1	1	0	0	1
	2	0	0	2
	3	0	0	4
	4	0	0	6
Tier 2	4	1	0	6
	4	2	0	8
	4	4	0	10
Tier 3	4	4	1	13
	4	4	2	16
	4	4	3	18
	4	4	4	20

- Medium-transfer problems (immediate and retention): These problems were scored in the same manner as near-transfer problems (Table 5) because the task was conceptually the same: solving the first layer. The only difference that creates the small transfer component is the color of the first layer.
- Far-transfer problems (immediate and retention): Participants were awarded points for every piece inserted correctly, regardless of whether it was an edge piece or corner piece (scoring scheme will not be tiered like in near- and medium-transfer problems); however, due to difficulty, each of the four edge pieces were given slightly more weight (3 points) than each of the four corner pieces (2 points). The maximum score for a far-transfer Rubik's Cube task was therefore 20 points.

CHAPTER 4. RESULTS AND DISCUSSION

The demographic averages for participants in Experiment 1 (cryptarithmetic) and Experiment 2 (Rubik's Cube) are shown in Table 7.

Table 7.

Demographic averages for Experiment 1 and Experiment 2 participants

Domain	Demographic	Mean	Standard deviation
Cryptarithmetic (Experiment 1)	Gender	56% female	
	Age	19.1	2.9
	Year in school	2.4	1.4
	Major	88% engin/sciences	
	GPA	3.39	0.55
	SAT	1390	121
	Expected difficulty (Likert 1-7)	4.33	1.04
Rubik's Cube (Experiment 2)	Gender	58% female	
	Age	19.5	3.1
	Year in school	2.3	1.3
	Major	92% engin/sciences	
	GPA	3.50	0.41
	SAT	1419	110
	Expected difficulty (Likert 1-7)	4.71	0.97

There were no significant differences in the participants between the two experiments, and furthermore, no systematic differences in pre-existing ability were found between the participants of any of the conditions within each domain, preventing pre-existing ability from being a confounding factor in the current experiments. Table 8 summarizes the results of the tests of pre-existing abilities.

Table 8.

Statistics from tests of pre-existing abilities

		Mean test score (SD)	Correlation
Experiment 1 (cryptarithmetic)	Ability 1 (eq. systems) Ability 2 (logic)	27.9% (18.9%) 44.9% (23.7%)	$r = 0.163$
Experiment 2 (Rubik's Cube)	Ability 1 (paper folds) Ability 2 (cube unfolding)	72.5% (14.9%) 68.5% (22.7%)	$r = 0.553$

Two ratings were provided for each of the four survey questions in which judgment was necessary and inter-reliability was analyzed. Those questions and their intraclass correlation coefficients (ICC) are listed here (the reliability was high, as the value of each coefficient exceeded 80%):

- "What knowledge gaps do you plan on filling during the next learning period?" (ICC of consistency = 0.835)
- "What do you think was the purpose of the problem-solving learning period (as opposed to the instructional learning period)?" (ICC of absolute agreement = 0.965)
- "Regarding the technical domain content, what potential learning mistakes by future participants would you like to warn them about?" (ICC of consistency = 0.853)
- "List all solution strategies that you have used so far (can be general problem-solving strategies or domain-specific methodologies)" (ICC of consistency = 0.832)

In the tables of results, statistically-significant findings ($p < 0.05$) are highlighted and findings with trending statistical significance ($p < 0.1$) are asterisked.

4.1 Instruction type main effects

4.1.1 Primary immediate learning assessments

A general linear model (GLM) was created to analyze how the manipulated independent variables and participant pre-existing ability affected immediate post-test scores in both domains. For each individual problem type as well as overall test score, the data indicated that there was no significant difference between productive failure and direct instruction, except for one instance (medium-transfer problem in the cryptarithmetic domain) that is likely a random outlier given the pattern of the other results. Table 9 outlines these results (maximum possible test score is 100%).

Table 9.

Post-test score differences between instruction types

Domain	Transfer type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						PF	DI
Cryptarithmetic	Near	2.55	699.427	0.127	0.118	92.3 (23.3)	75.9 (33.1)
	Medium	6.419	467.122	0.02	0.253	90.7 (23.4)	69.4 (34.1)
	Far	0.295	1247.56	0.594	0.015	59.4 (30.0)	66.8 (35.4)
	Total	1.865	343.95	0.188	0.089	81.4 (13.2)	71.6 (22.2)
Rubik's Cube	Near	1.167	927.376	0.294	0.058	48.0 (35.7)	61.2 (37.2)
	Medium	1.064	879.005	0.315	0.053	46.4 (29.5)	58.7 (35.1)
	Far	0.378	444.073	0.546	0.019	76.6 (19.9)	71.4 (30.3)
	Total	0.522	561.581	0.479	0.027	56.9 (25.1)	63.8 (32.3)

In the realm of near-transfer test problems, it was not expected that productive failure would produce significantly better task performance than direct instruction, especially when the problems were administered immediately after learning has occurred. This expectation was realized in the above results. Many of the hypothesized advantages of PF methods were expected to instead become manifest during medium- and far-transfer problems, as well as retention problems, while DI methods' usage of isomorphic problems as practice (Clark, Kirschner, &

Sweller, 2012) are conducive to performance on test problems that are similar to the practiced ones. The "regurgitative" nature of completing procedurally-similar problems immediately after learning increases the importance of streamlined problem-solving search processes often emphasized in DI (Rourke & Sweller, 2009) while rendering the potentially deeper structural learning in PF relatively less useful.

However, a reason that DI was not hypothesized to actually overtake PF in immediate near-transfer task performance is that PF participants tend to report greater curiosity during canonical instruction than DI participants do (Loibl & Rummel, 2014b), a phenomenon that was indirectly observed in this study when participants were surveyed about the purpose of the problem-solving learning period. In the cryptarithmetic domain, PF participants (M = 95%) were significantly more likely than DI participants (M = 24%) to say that the problem-solving period was to be used for exploration (as opposed to practice and application), $F(1, 43) = 43.711$, $MSE = 0.128$, $p = 0.000$, partial $\eta^2 = 0.504$ (mean difference = 71%); a similar pattern of results for PF (M = 100%) and DI (M = 30.8%) held in the cube domain, $F(1, 49) = 54.044$, $MSE = 0.113$, $p = 0.000$, partial $\eta^2 = 0.524$ (mean difference = 69.1%). This question served to illuminate the mindsets of participants in the two instructional conditions and indeed revealed the exploratory approaches that PF participants tended to take.

According to Loibl and Rummel (2014b), initial unguided problem-solving periods in PF help learners to identify knowledge gaps that they are then more curious about resolving later when canonical instructions are presented; DI learners are not given intrinsic reason to pay as much attention to the canonical instructions. The benefits of the extra attention paid by PF participants to canonical instructions should be particularly evident during near-transfer test problems, given that the instructions focus on those types of problems. Moreover, not only were

PF learners expected to be more curious and engaged, they were also expected to be more able to appreciate critical features of the presented canonical solutions due to comparisons of the strengths and weaknesses of their invented solutions and the canonical ones (Moore & Schwartz, 1998). Therefore, the advantages for each method were expected to "cancel out" to some extent, and the non-significant differences between PF and DI in both domains fulfilled those expectations.

Productive failure was hypothesized to produce significantly better performance in medium- and far-transfer problems, but that largely turned out not to be the case. The hypothesis was based on the notion that PF methods, just through the order of instruction, would require learners to combine heuristics and formal knowledge in ways that the "canonical instruction, then application practice" order in DI does not (Kapur & Bielaczyc, 2011). This combining of various knowledge bases in PF was expected to provide learners with the resources to generate relatively wide ranges of solution methods (diSessa & Sherin, 2000) due in part to the exploratory information gleaned from the initial problem-solving periods, and these different solution methods should have enabled better attempts at transfer problems that cannot be solved solely using canonical instructions. Participants in PF conditions (M = 0.594 unique solution strategies, SD = 0.837) did indeed attempt unique solution strategies more often than DI participants (M = 0.219, SD = 0.420) in cryptarithmetic, $F(1, 62) = 5.131$, $MSE = 0.439$, $p = 0.027$, partial $\eta^2 = 0.076$ (mean difference = 0.375), and the Rubik's Cube domain revealed similar differences between PF (M = 0.781, SD = 0.552) and DI (M = 0.375, SD = 0.492), $F(1, 62) = 9.648$, $MSE = 0.274$, $p = 0.003$, partial $\eta^2 = 0.135$ (mean difference = 0.406).

However, the use of unique strategies (those that were not explicitly explained in instructional material) apparently did not aid participants on tasks of medium and far transfer.

While it still might be the case that those tasks do require novel and creative solution methods, perhaps the participants' invented methods were either not particularly relevant or did not enable the participants to learn deep structural information about the domain. Furthermore, deciphering the parts of a solution attempt that are generalizable, and those that are context-specific and ungeneralizable, is often difficult for novices due to a lack of experience (Patel, Groen, & Norman, 1993), an issue that is likely magnified in PF when participants initially are relying more on their own heuristics to make assumptions about the domain. The participants in PF conditions might not have performed as well as expected on tasks of further transfer because they could not reliably discern how much to generalize from their invented solution attempts, whereas DI participants received more guidance on that front.

4.1.2 Primary retention learning assessments

To analyze the retention test performance dependent measure, the four independent variables, pre-existing ability (covariate), and immediate post-test score (covariate), were used as predictors in a GLM. No significant retention score differences were found between PF (M = 45.94%, SD = 21.62%) and DI (M = 48.62%, SD = 20.10%) in cryptarithmetic, $F(1, 18) = 0.114$, $MSE = 376.147$, $p = 0.739$, partial $\eta^2 = 0.006$ (mean difference = 2.68%), and no significant retention score differences were found between PF (M = 63.72%, SD = 26.6%) and DI (M = 66.35%, SD = 26.6%), in Rubik's Cube, $F(1, 16) = 0.219$, $MSE = 171.214$, $p = 0.646$, partial $\eta^2 = 0.014$ (mean difference = 2.63%).

It was hypothesized that the inherently frequent activation of prior and long-term knowledge during initial PF problem-solving would require learners to connect new material with relatively stable information that they already knew (Kapur, 2012) and furthermore lead to deeper encoding and assembling of schemas (Hiebert & Grouws, 2007). As a result, the learning

that ensued was expected to be more enduring and less fleeting, a difference that would be most apparent on retention problems. When surveyed on a Likert scale (1-7, 7 = most), participants in PF (M = 4.25, SD = 2.11) did not report using significantly more prior knowledge than DI (M = 4.03, SD = 1.56) in cryptarithmetic, $F(1, 62) = 0.223$, $MSE = 3.435$, $p = 0.639$, partial $\eta^2 = 0.004$ (mean difference = 0.22) and the differences between PF (M = 3.31, SD = 1.79) and DI (M = 3.13, SD = 1.66) were also not statistically significant in Rubik's Cube, $F(1, 62) = 0.189$, MSE = 2.974, $p = 0.665$, partial $\eta^2 = 0.003$ (mean difference = 0.188). For now, these data can inform some discussion and conclusions, but more-detailed analyses are likely needed in the future to examine, more generally, the differences in how PF and DI participants used problem-solving periods. Question prompts during problem-solving, for example, could enable researchers to more deeply study why a participant invented a particular solution strategy and whether that strategy contributed any generalizable domain knowledge through its use, or how a participant could be encouraged to activate more relevant prior and long-term knowledge.

In the current experiments, given that PF methods did not prove superior to DI in terms of forcing participants to lean more on their prior knowledge, it is then unsurprising that retention performance was about equal between the two conditions. This pattern of findings on retention performance contradicts what "desirable difficulties" research would predict (e.g., Bjork, 2013), if indeed productive failures are supposed to function like desirable difficulties (i.e., slow performance improvements early on due to difficulty designed into the instruction, but better performance later). If they are supposed to, it would be expected that PF participants surpass their DI counterparts on assessments like the retention test, which was administered one week after the material was learned. Participants' struggles during the PF generation period would require deeper and more durable processing to navigate (i.e., connected to prior

knowledge and/or self-generated heuristics), while DI participants would be more likely to fall into a false sense of competency because the learning process is relatively easier and performance on immediate tasks improves relatively quickly (Marsh & Butler, 2013). However, survey measures such as workload (via NASA TLX, the results of which will be detailed more in the next section) revealed that PF was not an appreciably more difficult experience than DI, and in some instances was actually reported to be an easier experience. Furthermore, not all participants in PF actually failed after the initial "struggle" period, which likely means that the given tasks were not difficult enough to yield productive failures and the associated benefits: 8 of 32 cryptarithmetic participants scored 100% on the mid-point check, while 6 of 32 Rubik's Cube participants performed likewise. Therefore, PF did not create enough desirable difficulty for participants, and as a result, retention performance was not improved.

Methodologically, the possibility exists that one week was not a long enough time period for retention differences between the instruction types to become manifest, although the statistics discussed previously regarding use of prior knowledge and reported difficulty imply that elongating the time still might not have revealed a difference. Instead, a future research direction might involve more explicit elicitation of prior/heuristic knowledge during PF generation periods, perhaps with scaffolding to ensure the domain relevance of that knowledge.

4.1.3 Mid-point check and secondary survey assessments

Between the first and second learning periods, all participants completed a mid-point progress check by attempting a near-transfer problem (an addition cryptarithmetic problem or first layer of Rubik's Cube, depending on assigned domain). The performance differences (maximum score of 100%) between PF (M = 40.63%, SD = 40.75%) and DI (M = 53.57%, SD = 38.78%) on this mid-point problem were non-significant for cryptarithmetic, $F(1, 62) = 1.695$,

MSE = 1582.299, $p = 0.198$, partial $\eta^2 = 0.027$ (mean difference = 12.94%), and the differences between PF ($M = 43.44\%$, $SD = 34.65\%$) and DI ($M = 33.28\%$, $SD = 22.95\%$) in Rubik's Cube were also not significant, $F(1, 62) = 1.911$, *MSE* = 1650.391, $p = 0.172$, partial $\eta^2 = 0.03$ (mean difference = 10.16%). This finding is surprising given that at the mid-point, DI participants are the only ones to have experienced canonical instruction and were therefore hypothesized to perform better on this problem. A near-transfer problem administered immediately after instruction is the type of problem that students using direct instruction should theoretically solve particularly well given the "learn, then apply" order of instruction to that point (Clark, Kirschner, & Sweller, 2012). However, in the current experiments, participants in PF did not report a significantly higher number of knowledge gaps after the first learning period than their counterparts in DI, suggesting that the generation period (PF) did not induce as much failure as intended; there was no significant difference between PF ($M = 0.56$ reported knowledge gaps, $SD = 0.67$) and DI ($M = 0.53$, $SD = 0.51$) for cryptarithmetic, (1, 62) = 0.044, *MSE* = 0.352, $p = 0.834$, partial $\eta^2 = 0.001$ (mean difference = 0.03), and no significant difference between PF ($M = 0.81$, $SD = 0.69$) and DI ($M = 0.59$, $SD = 0.56$) in Rubik's Cube, $F(1, 62) = 1.93$, *MSE* = 0.397, $p = 0.17$, partial $\eta^2 = 0.03$ (mean difference = 0.22).

Another surprising finding at the mid-point was the participants' reported workload via NASA TLX. Table 10 summarizes the workload statistics (maximum possible reported workload is 100%):

Table 10.

Workload differences between instruction types, mid-point (TLX)

Domain	Workload type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						PF	DI
Cryptarithmetic	Mental	2.608	381.821	0.112	0.042	55.5 (22.7)	63.6 (16.8)
	Temporal	2.870	566.27	0.095*	0.046	33.2 (25.3)	43.4 (23.2)
	Effort	4.747	438.503	0.033	0.073	47.7 (21.8)	59.1 (21.1)
	Frustration	1.075	768.607	0.304	0.018	38.8 (30.4)	46.0 (24.1)
Rubik's Cube	Mental	0.000	373.497	0.994	0.000	60.3 (17.9)	60.2 (20.9)
	Temporal	29.258	413.527	0.000	0.339	29.9 (17.6)	58.1 (22.4)
	Effort	1.779	374.592	0.188	0.018	63.6 (17.8)	57.0 (22.0)
	Frustration	0.085	731.136	0.772	0.001	48.5 (27.6)	50.7 (28.4)

Minimal guidance methods in the past have usually induced greater workload when compared to direct instruction (Hardiman, Pollatsek, & Weil, 1986), and PF, as a minimal guidance method, was hypothesized to be no different. The requisite learner engagement to freely explore a problem space (Durkin & Rittle-Johnson, 2012) was expected to require more mental resources than proceeding through comparatively straightforward canonical instruction. However, the data show that the workload differences between the instruction types were usually not significant, with the only exceptions being significant differences in the opposite of the expected direction: the two significant differences as indicated in Table 8, and a third trending significant difference for temporal workload in cryptarithmetic (DI producing heavier workload than PF). The fact that temporal workload was reported as higher (i.e., at least trending significance) in DI than PF for both domains is interesting and can perhaps be explained by the fact that DI participants were given a concrete amount of material to study during the first learning period and therefore felt pressure to read through all of the material before the end of the period. Conversely, PF participants were more likely to use the first learning period for exploration and therefore did not feel pressure to complete a concrete task, per se.

If PF methods were to reduce stressors such as working memory load, it could possibly be achieved through the increased use of prior long-term knowledge instead of working memory (Kapur & Bielaczyc, 2011). However, previously-reported data showed that participants in the current experiments accessed prior knowledge at roughly the same rates regardless of instruction type. Furthermore, questions remain as to whether hypothetically low workload would confer benefits to learners. Kapur (2014) has found instances in which higher mental workload can co-exist with better learning, and Vygotsky (1978) among others has hypothesized before that there might exist a "sweet spot" of mental workload that produces the best learning. Therefore, more research is needed to determine the nuances of the relationship between workload and learning.

4.1.4 Secondary post-learning survey assessments

After the second learning period, all participants completed a survey that provided more insight regarding the effects of instruction type. As described previously, PF methods were able to accomplish two things significantly better than DI methods: A) create an exploratory mindset for participants during problem-solving, and B) induce a wider range of unique solution attempts from participants during problem-solving. However, PF methods did not better facilitate participants reflecting on flaws in their mental models, compared to DI methods; when asked to identify their mistakes that future participants should be warned about, PF (M = 0.69 potential mistakes, SD = 0.69) and DI (M = 0.50, SD = 0.51) participants exhibited no significant differences in the number of responses for cryptarithmetic, $F(1, 62) = 1.525$, $MSE = 0.369$, $p = 0.222$, partial $\eta^2 = 0.024$ (mean difference = 0.19), and in Rubik's Cube, PF (M = 0.66, SD = 0.55) and DI (M = 0.41, SD = 0.56) participants also exhibited no significant differences, $F(1, 62) = 3.274$, $MSE = 0.305$, $p = 0.075$, partial $\eta^2 = 0.05$ (mean difference = 0.25). Therefore, while PF functioned to some extent in encouraging exploration and diverse solutions, it failed to

elicit the deep reflection and recognition of flawed understanding that is crucial for students to allocate attention to the most relevant material (Durkin & Rittle-Johnson, 2012) and, ultimately, to learn (Chi, 2000). Given that performance differences between the instruction types were non-significant across all domains and timings, it can be hypothesized that recognizing flaws in understanding, along with some other possible cognitive processes, is likely to be a key process that unlocks the full potential of productive failure (i.e., just encouraging exploration and diverse solutions is apparently not enough).

Like they did at the mid-point survey, participants rated their subjective workload levels at the end of the second learning period. Table 11 summarizes this workload data (maximum possible reported workload is 100%):

Table 11.

Workload differences between instruction types, post-learning (TLX)

Domain	Workload type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						PF	DI
Cryptarithmetic	Mental	2.253	1821.673	0.139	0.036	60.5 (20.8)	76.6 (57.3)
	Temporal	0.360	549.785	0.551	0.006	50.8 (25.3)	47.3 (23.4)
	Effort	4.213	328.210	0.044	0.066	55.1 (17.5)	64.4 (19.3)
	Frustration	0.029	665.143	0.866	0.000	38.4 (27.3)	39.5 (23.9)
Rubik's Cube	Mental	0.165	341.510	0.686	0.003	62.0 (19.0)	63.9 (18.2)
	Temporal	2.227	620.905	0.141	0.360	41.0 (25.2)	50.3 (24.3)
	Effort	0.023	514.108	0.880	0.000	60.9 (21.5)	60.0 (23.9)
	Frustration	0.188	788.613	0.670	0.003	52.0 (29.4)	55.0 (26.6)

The workload differences between PF and DI are even less at post-learning than at the mid-point, likely due in part to the fact that participants of both instructional conditions have gone through the same problem-solving period and canonical instruction period (albeit in opposite order from the other condition), thereby decreasing the variation in experience somewhat. This finding is corroborated by the non-significant differences in perceived difficulty

of material (Likert scale 1-7, 7 is most difficult): The learning experience in PF (M = 4.16, SD = 1.39) was not perceived to be more difficult than DI (M = 4.34, SD = 1.15) when studying cryptarithmetic, $F(1, 62) = 0.344$, $MSE = 0.563$, $p = 0.56$, partial $\eta^2 = 0.006$, and the differences between PF (M = 4.22, SD = 1.31) and DI (M = 4.78, SD = 1.26) were also not significant in the Rubik's Cube domain, $F(1, 62) = 3.049$, $MSE = 1.66$, $p = 0.086$, partial $\eta^2 = 0.047$. Prior research has suggested that higher workload can induce participants to report higher subjective difficulty (Reynolds & Caperton, 2011); a similar phenomenon was expected to occur in the present studies, but the workload measures from NASA TLX indicate that workload was no higher in productive failure than in direct instruction.

Interest in the material, as measured by two Likert survey questions (1-7, 7 indicating high interest), did not appear to correlate with the aforementioned workload and perceived difficulty measures. In the cryptarithmetic domain, PF participants (M = 5.30, SD = 1.02) reported being significantly more interested in the material than DI participants (M = 4.42, SD = 1.27), $F(1, 62) = 9.208$, $MSE = 1.33$, $p = 0.004$, partial $\eta^2 = 0.129$ (mean difference = 0.88). The fact that the cryptarithmetic domain functions much like algebra, and is therefore not novel to most college students, likely created conditions in which instructional design accounted for much of the variance in stimulating learners.

In the Rubik's Cube domain, the difference in interest between PF (M = 4.91, SD = 1.33) and DI (M = 4.52, SD = 1.12) in the Rubik's Cube domain was non-significant, $F(1, 62) = 1.62$, $MSE = 1.507$, $p = 0.208$, partial $\eta^2 = 0.025$ (mean difference = 0.39). The Rubik's Cube, a domain that is unlike most traditional school subjects, likely presented tasks that were inherently interesting and novel to learners, independent of the instructional method used. The curiosity

naturally induced in learners through PF methods (Loibl & Rummel, 2014b) is not important when the domain itself is stimulating.

Attributing this pattern of results to any given dimension of the domains is difficult given that the domains differ along several dimensions (as shown in Table 2), but relative familiarity stands out as perhaps one of the most plausible explanations. Therefore, in the practical sense of implementing productive failure in classrooms, the novelty and familiarity of the domain should be considered, and the intrinsic motivation levels of the students might also be a factor. In future research, systematically manipulating the relative familiarities of domains, and controlling on all other dimensions, would enable researchers to test this explanation more incisively.

4.2 Subgoal label main effects

4.2.1 Primary immediate and retention learning assessments

Upon examining the subgoal predictor of the GLMs for immediate test and retention test performance, a pattern emerged regarding scores across domains. Table 12 summarizes the scores of participants who received subgoals (SUB) and those who received non-labeled (NL) instructions (maximum possible test score is 100%):

Table 12.

Test score differences between subgoal- (SUB) and non-labeled (NL) instructions

Domain	Test timing	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						SUB	NL
Cryptarithmetic	Immediate	0.053	343.954	0.821	0.003	77.3 (18.7)	75.7 (18.5)
	Retention	0.002	376.147	0.968	0.000	42.7 (14.0)	44.0 (18.7)
Rubik's Cube	Immediate	3.659	561.581	0.071*	0.161	69.2 (29.4)	51.4 (26.0)
	Retention	4.543	484.793	0.040	0.109	68.5 (26.3)	55.2 (27.7)

In the cryptarithmetic domain, subgoal labels appeared to make very little difference in test scores. Previous research has demonstrated that subgoal labels outline high-level information

that can help learners organize domain content in meaningful ways (Atkinson et al., 2000), which theoretically should improve performance. However, it is probable that the college-educated participants did not require subgoal labels to help them organize content in a domain that is similar to algebra.

According to the data, Rubik's Cube participants were aided greatly by subgoal labels. Sweller (2010) notes that subgoals enable learners to focus just on fundamental structures of problems and not incidental features. In a domain like the Rubik's Cube in which participants likely do not possess much relevant experience, this generalizable information from subgoal labels is crucial so that participants do not extrapolate from concepts that might have been specific only to a given example.

4.2.2 Workload measures (NASA TLX)

Some evidence suggests that the subgoal labels in cryptarithmetic, if anything, served only to increase participant workload, possibly because of extra effort needed to interact with them. Tables 13 and 14 outline the workload data for both domains (maximum possible reported workload is 100%).

Table 13.

Cryptarithmetic: Workload differences between subgoal-labeled and non-labeled instructions

Timing	Workload type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						SUB	NL
Mid-point	Mental	4.388	381.921	0.040	0.068	64.5 (19.1)	54.3 (20.4)
	Temporal	2.960	566.270	0.091*	0.047	43.4 (24.1)	33.2 (24.4)
	Effort	2.977	438.503	0.089*	0.048	58.0 (21.8)	48.9 (21.6)
	Frustration	0.269	768.607	0.606	0.004	44.2 (29.5)	40.6 (25.6)
Post-learning	Mental	1.956	1821.673	0.167	0.032	76.0 (57.6)	61.1 (20.4)
	Temporal	4.604	549.785	0.036	0.071	55.3 (26.1)	42.7 (20.8)
	Effort	3.035	328.210	0.087*	0.048	63.7 (15.8)	55.8 (21.0)
	Frustration	1.243	665.143	0.269	0.020	42.5 (27.7)	35.3 (22.8)

Table 14.

Rubik's Cube: Workload differences between subgoal-labeled and non-labeled instructions

Timing	Workload type	*F*	*MSE*	*p*	partial η^2	Mean (SD)	
						SUB	NL
Mid-point	Mental	0.137	373.497	0.712	0.002	59.3 (17.7)	61.3 (21.1)
	Temporal	0.163	413.527	0.688	0.003	42.5 (21.7)	46.1 (27.5)
	Effort	1.038	388.965	0.313	0.018	57.8 (18.8)	62.8 (21.5)
	Frustration	0.269	768.607	0.606	0.004	44.9 (27.0)	54.5 (28.1)
Post-learning	Mental	0.371	341.510	0.545	0.006	61.6 (16.5)	64.4 (20.5)
	Temporal	0.476	620.905	0.493	0.008	43.5 (23.3)	47.8 (26.9)
	Effort	1.505	514.108	0.225	0.024	57.0 (21.2)	63.9 (23.6)
	Frustration	0.229	788.613	0.634	0.004	51.8 (25.6)	55.1 (30.3)

According to Table 13, subgoals increased workload significantly in the cryptarithmetic domain. Furthermore, subgoal labels did not improve performance in cryptarithmetic, suggesting that the increased load might have been extraneous. As was stated before, it is perhaps the case that subgoal labels were not necessary in the cryptarithmetic domain due to participants' familiarity with algebra, which could explain why participants reported subgoals as relatively taxing to interact with.

Subgoals did not increase workload in the Rubik's Cube domain, as demonstrated in Table 14. The participants likely found the Rubik's Cube subgoal labels to be essential information and therefore did not perceive them as difficult to engage. After all, the subgoal labels improved Rubik's Cube performance substantially (Table 12).

Given the relatively robust findings in previous research regarding how subgoals reduce cognitive load in learners (e.g., Renkl & Atkinson, 2002; Morrison, Margulieux, & Guzdial, 2015), the findings in the current experiments are surprising. In future experiments, methods of implementing subgoal labels (e.g., frequency of labeling, type of content conveyed, learner role

in generation of labels) could be manipulated to examine whether workload and performance results depend on the method of labeling.

4.3 Interaction between instruction type and presence of subgoal labels

Before the experiments started, it was hypothesized that subgoal-related gains would be more pronounced in PF conditions than in DI conditions. Subgoal labels presented during the PF generation period were expected to mitigate the chances that learners aimlessly pursued irrelevant objectives and formed structural misconceptions, risks that are inherent in any minimally-guided method (Brown & Campione, 1994). While subgoal labels are generally important in DI materials as well, they were expected to be relatively less so because DI participants received instruction at the start of the learning process that was at least somewhat organized whether subgoals were labeled or not, and the participants were merely applying learned knowledge during the problem-solving phase (Clark, Kirschner, & Sweller, 2012), likely using the subgoal labels just as reminders.

The data suggested that no such interaction between instruction type and subgoal labeling occurred during the experiments, regardless of domain or timing of test. Table 15 summarizes the statistics regarding the interactions.

Table 15.

Interaction between instruction type and subgoal labeling, immediate and retention test scores

Domain	Test timing	*F*	*MSE*	*p*	partial η^2	Significance
Cryptarithmetic	Immediate	0.290	343.954	0.596	0.015	NS
	Retention	1.128	376.147	0.302	0.059	NS
Rubik's Cube	Immediate	0.091	561.581	0.766	0.005	NS
	Retention	0.552	171.214	0.468	0.033	NS

Instead, a plausible explanation is that the positive effects of subgoals are relatively robust across various methods of instruction, but not necessarily across all domains (per findings

described in section 4.2.1). After all, the key purpose of subgoal labels is helping learners recognize fundamental components of a domain or problem space (Catrambone, 1998), a useful aid regardless of whether a learner is using productive failure or direct instruction (the particular hypothesized benefits for PF, described earlier in this section justifying the interaction hypothesis, are secondary and perhaps not as reliable). However, the extent to which that aid increases performance significantly might depend on the relative familiarity of the domain and how easily learners can discern fundamental components on their own in that given domain.

In summary, subgoal labels improved performance in the Rubik's Cube domain, regardless of instruction type, but failed to improve performance in the cryptarithmetic domain (also regardless of instruction type). A potential future research direction could involve manipulating the scaffolding mechanism used in PF instruction to examine whether other scaffolding mechanisms are more reliable across domains (e.g., self-explanation prompts, social discourse; Lin, Hmelo, Kinzer, & Secules, 1999). Preventing learners from failing unproductively and veering too far off track is a scaffolding mechanism that has been shown to be effective in general (e.g., training wheels; Carroll & Carrithers, 1984), but other methods could prove superior in particular learning contexts. A systematic examination of domains is also necessary to study how these various scaffolding mechanisms interact with domains of particular characteristics; for example, the motivational aspects of group discourse (Lin, Hmelo, Kinzer, & Secules, 1999) could improve learning relatively substantially in inherently uninteresting domains, but not spur much improvement in domains that are inherently more interesting.

4.4 Testing effect and its interactions with instruction type

The literature supporting the testing effect is robust (e.g., review by Eisenkraemer, Jaeger, & Stein, 2013), especially with regards to long-term retention, and it was therefore

expected that those receiving a post-test would outperform, on retention tests one week later, those who merely re-studied. An example of the robustness of the testing effect from the current experiments was the finding that the effect on retention of completing a post-test did not change depending on instruction type; the interaction between post-test presence and instruction type was non-significant for cryptarithmetic, $F(1, 40) = 0.046$, $MSE = 257.918$, $p = 0.832$, partial $\eta^2 = 0.001$, as well as Rubik's Cube, $F(1, 40) = 1.754$, $MSE = 484.793$, $p = 0.193$, partial $\eta^2 = 0.045$. However, the occurrence of this effect did depend on domain: Participants receiving a post-test ($M = 47.8\%$, $SD = 20.5\%$) indeed significantly outperformed their re-studying counterparts ($M = 38.9\%$, $SD = 11.4\%$) in the Rubik's Cube domain, $F(1, 40) = 4.194$, $MSE = 257.918$, $p = 0.047$, partial $\eta^2 = 0.095$ (mean difference = 8.9%), but the difference in retention scores between post-test ($M = 62.5\%$, $SD = 26.1\%$) and re-study conditions ($M = 61.2\%$, $SD = 28.0\%$) was non-significant for the cryptarithmetic domain, $F(1, 37) = 0.042$, $MSE = 484.793$, $p = 0.839$, partial $\eta^2 = 0.001$ (mean difference = 1.3%). One explanation of the testing effect is that learners often activate related surrounding concepts when attempting to retrieve a target concept from memory (e.g., during a post-test), thereby increasing the number of semantic pathways available for future reaching of that target concept in a way that re-studying does not (Collins & Quillian, 1972; Carpenter, 2009).

As described before, cryptarithmetic is a domain that functions much like mathematics-related subjects that the participants have studied before, and it can therefore be expected that many semantic pathways to cryptarithmetic target concepts are already formed and used during the learning process, regardless of whether the participants are pushed to activate surrounding concepts through a post-test. The fact that cryptarithmetic lends itself easily to permanent

external memories of work (e.g., calculations worked out on scratch paper) also increases the effectiveness of re-studying.

In the novel Rubik's Cube domain, participants are likely not very able to activate surrounding concepts when re-studying, and the post-test therefore provides an important push to do so through the act of retrieval. Furthermore, Rubik's Cube participants are not provided with natural records of their work, which impedes the effectiveness of re-studying; that is, the moves made on the way to a solution are not stored in any permanent or external fashion, and they are therefore not easily available for review. Participants must remember their moves or find a way to transcribe them, both of which are difficult.

This theory of "spreading activation" can account for the difference in results between the two domains, although the usual caveats apply regarding the several dimensions on which the domains differ. Any given dimension could be hypothesized as the most sensible reason for the empirical pattern of results, but the other dimensions are possibly confounding variables whose effects are not known.

4.5 Interaction between instruction type and time constraint

One of the key tenets of productive failure methods is that exploration is crucial for people to learn because it can induce failure and failure-related benefits. Given that relatively unconstrained time periods are most conducive to encouraging effective exploration (Kehoe, Stasko, & Taylor, 2001), it was hypothesized that the learning gains made by PF participants over their DI counterparts would be most pronounced in extended-time conditions and less pronounced in limited-time conditions. However, this expected interaction between instruction type and time constraint did not occur, according to the data in Table 16.

Table 16.

Interaction between instruction type and time constraint, immediate and retention test scores

Domain	Test timing	*F*	*MSE*	*p*	partial η^2	Significance
Cryptarithmetic	Immediate	0.070	24.240	0.794	0.004	NS
	Retention	4.079	257.918	0.052	0.087	NS
Rubik's Cube	Immediate	0.070	4.074	0.933	0.000	NS
	Retention	0.001	171.214	0.976	0.000	NS

One possible explanation for the lack of interaction is that although learners are often spurred to explore and spend additional study time when they perceive material to be relatively difficult (LaPorte & Nath, 1976), the subjective difficulty of PF was not reported as higher than DI in either domain (section 4.1.4). Some past research has reported higher perceived difficulty by PF participants (e.g., Reynolds & Caperton, 2011; Kapur, 2014), but such a phenomenon did not occur in the current experiments. If the participants in PF conditions did not perceive the material to be relatively more difficult, then perhaps the extra time did not benefit them any more than it did DI participants (i.e., extended time is most needed when material is difficult). This explanation is supported by the finding that there was also no significant difference between PF and DI participants in terms of the number of identified knowledge gaps they reported wanting to investigate (section 4.1.3), which suggests that participants in both conditions needed the extended time to roughly the same extent.

CHAPTER 5. CONCLUSIONS

Many of the primary hypotheses were not supported by the data, but some secondary findings emerged that could illuminate a path forward in future research. Table 15 lists each tested hypothesis and how the data did or did not support it.

Table 15.

Summary of outcomes in the present studies

Number	Description	Notes on outcome
H1	Medium- and far-transfer problems, immediate post-test: PF participants will score more highly than DI participants (no difference for near-transfer problems)	Not supported; differences were non-significant for all problem types
H2	Retention test problems: PF participants will score more highly than DI participants	Not supported; differences were non-significant
H3	Number of identified knowledge gaps, mid-point check: PF participants will identify more gaps than DI participants	Not supported; differences were non-significant
H4	Mental workload (TLX), mid-point check: PF participants will report higher workload than DI participants	Some evidence that DI induced higher workload than PF, but differences were generally non-significant
H5	Amount of prior knowledge and intuition used, mid-point check: PF participants will list more concepts than DI participants	Not supported; differences were non-significant
H6	Near-transfer problem, mid-point check: PF participants will score lower than DI participants	Not supported; differences were non-significant
H7	Role of problem-solving in learning process, post-learning: PF participants will be more likely to report that problem-solving was used for exploration while DI participants will be more likely to report that it was used for practice/application	Supported

Table 15 continued

H8	Identifying potential mistakes of future participants, post-learning: PF participants will identify more potential mistakes than DI participants	Not supported; differences were non-significant
H9	Number of unique solution strategies invented, post-learning: PF participants will use more unique solution strategies than DI participants	Supported
H10	Perceived difficulty of material, post-learning: PF participants' subjective levels of difficulty reported will be higher than those reported by DI participants	Not supported; differences were non-significant
H11	Mental workload (TLX), post-learning: PF participants will report higher workload than DI participants	Not supported; differences were non-significant
H12	All problem types: Participants receiving subgoals will score more highly than participants without subgoals	Supported in Rubik's Cube domain, but not in cryptarithmetic domain
H13	Mental workload (TLX), mid-point check and post-learning: Participants receiving subgoals will report lower subjective workload than participants without subgoals	Not supported; workload sometimes heavier with subgoals in cryptarithmetic, but no differences found in Rubik's Cube
H14	All problem types: Subgoals will improve performance for PF participants more than they improve performance for DI participants	No such interaction was found
H15	Testing effect: Presence of immediate post-test will increase retention scores of PF participants more than those of DI participants	No interaction was found between presence of post-test and instruction type; however, testing effect did occur in Rubik's Cube and not cryptarithmetic
H16	Time constraints: Performance improvements produced by extended time will be larger for PF participants than for DI participants	No such interaction was found

In general, PF methods in the present studies produced some ostensibly positive ancillary developments for learners (exploratory mindsets, diverse solution attempts, and occasionally lower workload). However, those ancillary developments did not lead to the ultimate goal of increasing post-test and retention test performance. This phenomenon suggests questions for further study such as whether the relevance and quality of learners' solution attempts should be regulated somehow (perhaps through the use of scaffolding methods other than subgoal labels), or whether lower workload is beneficial in this context or domains.

More work is also needed to clarify the relationship between learners' pre-existing abilities and instruction type: Past research indicates that high-ability learners tend to perform well in low-structure environments because of their ability to connect new information with prior knowledge (Peterson, 1987) and low-ability learners need higher amounts of structure because they are not as able to develop their own strategies (Snow, 1982). High ability and prior knowledge also widen the difficulty range of tasks that learners are willing to engage with (zone of tolerable problematicity; Elshout, 1985), which could further inform PF implementation if initial problem-solving periods prove to be difficult for learners. Though the regression analyses from the present studies revealed no significant interaction between instruction type and pre-existing ability in either domain, the relatively small sample sizes limited the potential for those interactions to be revealed. Furthermore, in the cryptarithmetic experiment specifically, correlations between pre-existing ability scores and test performance were generally weak, indicating that the tested abilities were not the dominant determining characteristics of task performance in this domain. To the extent that an interaction between instruction type and pre-existing ability actually exists in the cryptarithmetic domain, the suboptimal selection of ability tests is a plausible explanation as to why the interaction was not revealed in this experiment.

SAT Math scores did exhibit a relatively strong correlation, $r = 0.452$, with retention performance in cryptarithmetic; one potentially interesting finding was that the difference between "ability-immediate" correlations and "ability-retention" correlations was not statistically significant in either domain (when using SAT Math for cryptarithmetic and spatial ability for Rubik's Cube), indicating that higher-ability participants did not retain more than lower-ability participants over and above the higher scores expected of higher-ability participants. Clarifying the relationship between ability and performance, including the identification of relevant pre-existing abilities, is crucial to effective implementation of PF methods in classrooms where students could possess varying levels of pre-existing knowledge and abilities.

Research in productive failure is still in its early stages and therefore much work remains to be done in improving the method itself. Potential improvements include explicit elicitation of prior domain knowledge, more meaningful subgoal labels, and group learning implementation. Replicating findings in various domains will also be an important task for the future, given that people have access to (and interests in) learning wider varieties of information than ever but most learning research still centers on just science- and mathematics-related domains. Some patterns of results from the current experiments changed depending on domain, but systematic selection of domains would enable researchers to find more precisely the dimensions and characteristics of domains that drive changes in results (e.g., an experiment in which the two domains are equal on every dimension except for one domain having a spatial component). Productive failure has already shown the potential to change the way researchers, educators, and learners think about "the assistance dilemma," but there is much more to do.

APPENDIX A. DEMOGRAPHICS QUESTIONNAIRE

Participant number: ____

Demographics questionnaire

1. Gender (circle one): Male Female

2. Age: _____

3. Major: _____

4. Year in college: First Second Third Fourth Fifth Other: _____

5. College GPA (if you remember): _____ / 4.0

6. SAT scores (if you remember): Math _____ Verbal _____

7. Have you systematically learned how to solve the Rubik's Cube before ("systematically" means looking it up online and/or learning from a friend, not just messing around)? YES NO

8. How difficult do you expect learning to solve a Rubik's Cube will be (1 = very easy, 7 = nearly impossible)?

1 2 3 4 5 6 7

9. In cryptarithmetic, solved arithmetic problems are presented to you and your task is to find the value of each variable. A famous example of such a problem is shown to the right.

	S	E	N	D
+	M	O	R	E
M	O	N	E	Y

a. Have you systematically learned how to solve cryptarithmetic problems before? YES NO

b. How difficult do you expect learning to solve these types of problems will be (1 = very easy, 7 = nearly impossible)?

1 2 3 4 5 6 7

10. When was the last time you regularly used algebra? Report your answer in terms of years ago, and feel free to use decimals if need be (write "0" if currently in use regularly): Years ago: _____

APPENDIX B. INTRODUCTION TO CRYPTARITHMETIC

Introduction to cryptarithmetic
(adapted from ElitmusZone.com)

Cryptarithmetic is the science and art of creating and solving cryptarithms, a genre of mathematical puzzle in which the digits are replaced by letters of the alphabet or other symbols.

The world's best known cryptarithm is undoubtedly this one:

	S	E	N	D
+	M	O	R	E
M	O	N	E	Y

It was created by H.E. Dudeney and first published in the July 1924 issue of Strand Magazine associated with the story of a kidnapper's ransom demand.

The fundamental rules of cryptarithmetic are as follows:

1. Each variable should have **unique** and **distinct** value
2. Each letter represents only **one digit** throughout the problem
3. Numbers must not begin with zero (i.e., 0123 is wrong, although 123 could be correct)
4. Your task is to find the value of each letter in the cryptarithm
5. There must be **only one solution** to the problem
6. The numerical base, unless specifically stated, is 10
7. After replacing letters by their digits, the resulting arithmetic operations must be correct.

Your task today is to learn how to solve cryptarithmetic problems. During the learning process, you will have access to a cryptarithmetic tutorial and time allocated for problem-solving. After the learning process, you will be tested on your ability to solve various cryptarithmetic problems.

APPENDIX C. CRYPTARITHMETIC TUTORIAL EXAMPLE PAGES

C.1 With subgoals labeled

HOW TO SOLVE AN ADDITION CRYPTARITHM

1. Preparation

Rewrite the problem, expanding the space between the numbers to make room for trial numbers that will be written around the letters. For example, the puzzle SEND + MORE = MONEY should be written:

```
      S   E   N   D
+     M   O   R   E
-------------------
  M   O   N   E   Y
```

2. Search for 0s and 9s

A good hint to find 0s or 9s is to look for columns containing two or three identical letters.

Example 1:

```
      *   *   *   A
+     *   *   *   A
-------------------
      *   *   *   A
```

Example 2:

```
      *   *   *   B
+     *   *   *   A
-------------------
      *   *   *   B
```

In both of these examples, it follows that A = 0 (the property here is called the "additive property of zero," which states that adding zero to any number leaves that number unchanged).

Now look at these examples:

Example 3:

```
      *   A   *   *
+     *   A   *   *
-------------------
      *   A   *   *
```

Example 4:

```
      *   B   *   *
+     *   A   *   *
-------------------
      *   B   *   *
```

When the same letters are in the body of the cryptarithm, A can be 0 or 9, depending on whether a 1 was carried over from the previous column (9s mimic 0s every time there's a "carry 1").

3. Search for 1s

Look for single left-hand digits in the sum – every time adding N-digit numbers creates a sum of N+1 digits, the left-hand digit is 1. For example, M must be 1 here.

```
      S   E   N   D
+     M   O   R   E
-------------------
  M   O   N   E   Y
```

C.2 Without subgoal labels

HOW TO SOLVE AN ADDITION CRYPTARITHM

Rewrite the problem, expanding the space between the numbers to make room for trial numbers that will be written around the letters. For example, the puzzle SEND + MORE = MONEY should be written:

```
   S  E  N  D
+  M  O  R  E
-------------
M  O  N  E  Y
```

A good hint to find 0s or 9s is to look for columns containing two or three identical letters.

Example 1:

```
   *  *  *  A
+  *  *  *  A
-------------
   *  *  *  A
```

Example 2:

```
   *  *  *  B
+  *  *  *  A
-------------
   *  *  *  B
```

In both of these examples, it follows that A = 0 (the property here is called the "additive property of zero," which states that adding zero to any number leaves that number unchanged).

Now look at these examples:

Example 3:

```
   *  A  *  *
+  *  A  *  *
-------------
   *  A  *  *
```

Example 4:

```
   *  B  *  *
+  *  A  *  *
-------------
   *  B  *  *
```

When the same letters are in the body of the cryptarithm, A can be 0 or 9, depending on whether a 1 was carried over from the previous column (9s mimic 0s every time there's a "carry 1").

Look for single left-hand digits in the sum – every time adding N-digit numbers creates a sum of N+1 digits, the left-hand digit is 1. For example, M must be 1 here.

```
   S  E  N  D
+  M  O  R  E
-------------
M  O  N  E  Y
```

APPENDIX D. SUBGOAL OUTLINE FOR COMPLETING THE PRIMARY CRYPTARITHMETIC TASK (ADDITION PROBLEM)

Cryptarithmetic addition subgoals

1. Write the problem such that there's enough space to write trial numbers around each letter

2. Search for 0s and 9s (adding 0 to any number leaves that number unchanged)

3. Search for 1s (i.e., in some situations, the left-most digit of the sum will have to be a 1 due to limits on how large the sum can feasibly be)

4. Generate possible values for variables and test them, using systems of equations and logic puzzle reasoning to find the values for each variable

APPENDIX E. MID-POINT QUESTIONNAIRE

Mid-point questionnaire

What percentage of an example problem do you think you would be able to complete right now?

Percentage: _____

What knowledge gaps do you plan on filling during the next learning period (it's okay if there aren't any)?

To what extent did this domain content remind you of other tasks you have done before (1 = not at all, 7 = have done something almost just like it)?

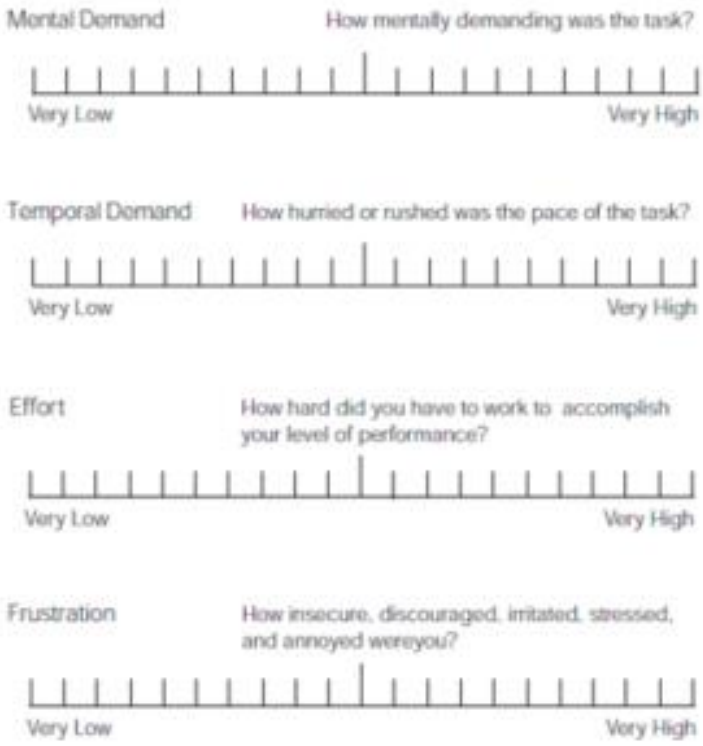
1 2 3 4 5 6 7

How much has your intuition and prior knowledge helped with the task so far (1 = not at all, 7 = indispensable)?

1 2 3 4 5 6 7

(MORE QUESTIONS ON BACK)

Draw an "X" on the line that most represents your level of workload on each of the following scales:



APPENDIX F. POST-LEARNING QUESTIONNAIRE

Post-learning questionnaire

What do you think was the purpose of the problem-solving learning period (as opposed to the instructional learning period)?

Regarding the technical domain content, what potential learning mistakes by future participants would you like to warn them about?

List all solution strategies that you have used so far (can be general problem-solving strategies or domain-specific methodologies):

How difficult did you find this material (1 = very easy, 7 = nearly impossible)?

1 2 3 4 5 6 7

What percentage of the given task do you think you can execute at this point?

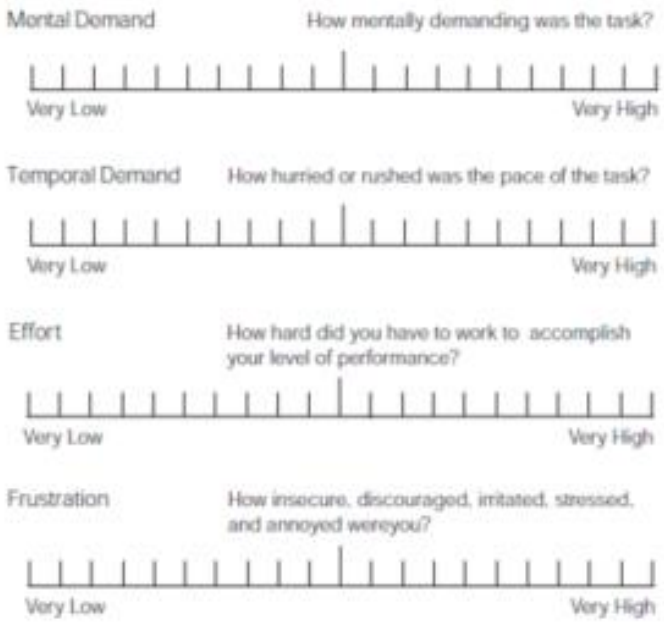
Percentage: _____

What percentage of the given task do you think you could execute **when you come back in a week**?

Percentage: _____

(MORE QUESTIONS ON BACK)

Draw an "X" on the line that most represents your level of workload on each of the following scales:



Please rate the following statements on a 7-point scale
(1 = not at all true of me, 7 = very true of me)

A1. I am interested in the material that I'm learning in this study

1 2 3 4 5 6 7

A2. I'm learning some things that will be useful to me in some way

1 2 3 4 5 6 7

B. I was surprised or confused, in any way, upon finding out the solutions to the presented problems

1 2 3 4 5 6 7

APPENDIX G. CRYPTARITHMETIC DOMAIN IMMEDIATE POST-TEST DIRECTIONS

Cryptarithmetic post-test

- Use scratch paper to do your work and report just your answers on this sheet.
- If you can't figure out a letter, you can put two possible values down (you'll receive half-credit if the correct value is one of those two values)

Problem 1 (5 min): Solve the following addition cryptarithm

```
  E A T
+ T H A T
---------
A P P L E
```

	Value 1	Value 2 (if needed)
A		
E		
H		
L		
P		
T		

Problem 2 (6 min): Solve the following addition cryptarithm

```
    N O
  G U N
+   N O
---------
H U N T
```

	Value 1	Value 2 (if needed)
G		
H		
N		
O		
T		
U		

Problem 3 (9 min): Solve the following subtraction cryptarithm

```
  F O O L
-   E L F
---------
    E L F
```

	Value 1	Value 2 (if needed)
E		
F		
L		
O		

APPENDIX H. CRYPTARITHMETIC DOMAIN RETENTION TEST DIRECTIONS

Cryptarithmetic retention test

- Use scratch paper to do your work and report just your answers on this sheet.
- If you can't figure out a letter, you can put two possible values down (you'll receive half-credit if the correct value is one of those two values)

Problem 1 (5 min): Solve the following addition cryptarithm

```
  B A S E
+ B A L L
---------
G A M E S
```

	Value 1	Value 2 (if needed)
A		
B		
E		
G		
L		
M		
S		

Problem 2 (6 min): Solve the following addition cryptarithm

```
  T A K E
      A
+ C A K E
---------
  K A T E
```

	Value 1	Value 2 (if needed)
A		
C		
E		
K		
T		

Problem 3 (9 min): Solve the following subtraction cryptarithm

```
  C O U N T
-   C O I N
-----------
    S N U B
```

	Value 1	Value 2 (if needed)
B		
C		
I		
N		
O		
S		
T		
U		

APPENDIX I. CRYPTARITHMETIC DOMAIN TESTS OF PRE-EXISTING ALGEBRAIC AND LOGIC ABILITY

I.1 Examples of equation system problems

Solve each system by substitution.

1) $y = 6x - 11$
$-2x - 3y = -7$

2) $2x - 3y = -1$
$y = x - 1$

I.2 Example of logic puzzle and question

A panel of music historians ranked eight contemporary songwriters – Jackson, King, Lennon, Mitchell, Nicks, Prince, Simon, and Wonder – according to their relative impact on the evolution of the popular song form. No other songwriters were considered, and there were no ties in the final ranking. The ranking of the songwriters met the following conditions:

Nicks was ranked higher than Lennon but lower than Simon.

Prince was ranked lower than both Mitchell and Jackson.

Wonder was ranked lower than Nicks.

Jackson was ranked higher than Simon.

Nicks was ranked higher than King.

1. Which one of the following could represent the ranking of songwriters, listed from highest to lowest?

- (A) Jackson, Simon, King, Mitchell, Prince, Nicks, Lennon, Wonder
- (B) Jackson, Simon, Prince, Nicks, Mitchell, Wonder, Lennon, King
- (C) Mitchell, Simon, Jackson, Prince, Nicks, Lennon, Wonder, King
- (D) Mitchell, Jackson, Simon, Nicks, King, Wonder, Lennon, Prince
- (E) Mitchell, Jackson, Prince, Simon, Lennon, Wonder, Nicks, King

APPENDIX J. INTRODUCTION TO RUBIK'S CUBE AND TUTORIAL

Introduction to Rubik's Cube and tutorial interface

The Rubik's Cube is a 3x3x3 cube composed of twenty-six visible pieces:

- Six center pieces,
- twelve edge pieces,
- eight corner pieces,

and one center spindle. The types of visible pieces are described below (look at your cube to examine the properties of each of these pieces):

- *Center pieces*: The most important step to understanding the cube is realizing that the center pieces are fixed (i.e. they do not move). Therefore, the relative positions of the center pieces are also fixed (e.g. center pieces on opposite sides of each other will always stay that way – white and yellow, for example). Each of the center pieces is located at the center of a face and each center piece has a distinct color (one of the six different colors of the cube). Note: A side is defined by the color of its center piece – e.g. the "yellow side" is the side with the yellow center piece.
- *Edge pieces*: Each edge piece has two colors and these two colors always represent colors from adjacent sides (e.g. in Figure 1, the red and white sides, which are next to each other); therefore, an edge piece can never have two colors that are represented by opposite sides, such as white and yellow (the white and yellow center pieces are on opposite sides of the cube).
- *Corner pieces*: Each corner piece has three colors and these three colors always represent sides that meet at one point (e.g. in Figure 1, the colors red, green, and white). As with edge pieces, opposite colors cannot be represented on one corner piece (e.g. red and orange – verify by looking at your cube).

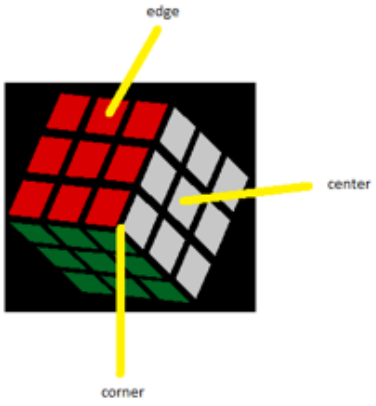


Figure 1. The visible pieces of a Rubik's Cube

In general, the cube is manipulated through the turning of its faces. Each of the faces moves independently of the other faces and can be turned through 360 degrees of rotation. When the Rubik's Cube is fully solved, each face consists of only one color.

The first step in solving a Rubik's Cube is to create "the cross," which involves placing four edge pieces around a center piece, all of the same color (see Figure 2) – in addition, the four edge pieces must be oriented such that each side of the edge piece is adjacent to a center piece of the same color (for example, the red/yellow edge piece is oriented such that the yellow side is adjacent to the yellow center and the red side is adjacent to the red center). By convention, most people solve the cross on the white side (the side with a white center piece), but the cross can be made on any side. Figure 2 below shows what a completed cross looks like if done from the yellow side.

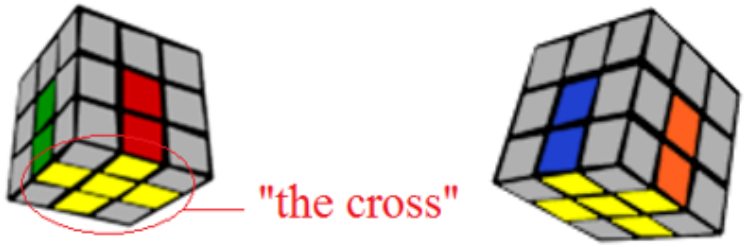


Figure 2. Completed yellow cross

The next step to solving the Rubik's Cube is to complete the first layer; to do that, you will have to place the four yellow corner pieces into the slots around the cross such that the two non-yellow colors of the corner pieces match the colors of the edge pieces that they're adjacent to. Figure 3 below shows a cube with a fully completed first layer.

Figure 3. Completed first layer

Your task today is to learn how to solve the first layer. You will have access to a tutorial that will include videos resembling the interface shown in Figure 4, and you will also have time allocated for problem-solving. After finishing the learning process, you will be tested on your ability to solve various Rubik's Cube problems.

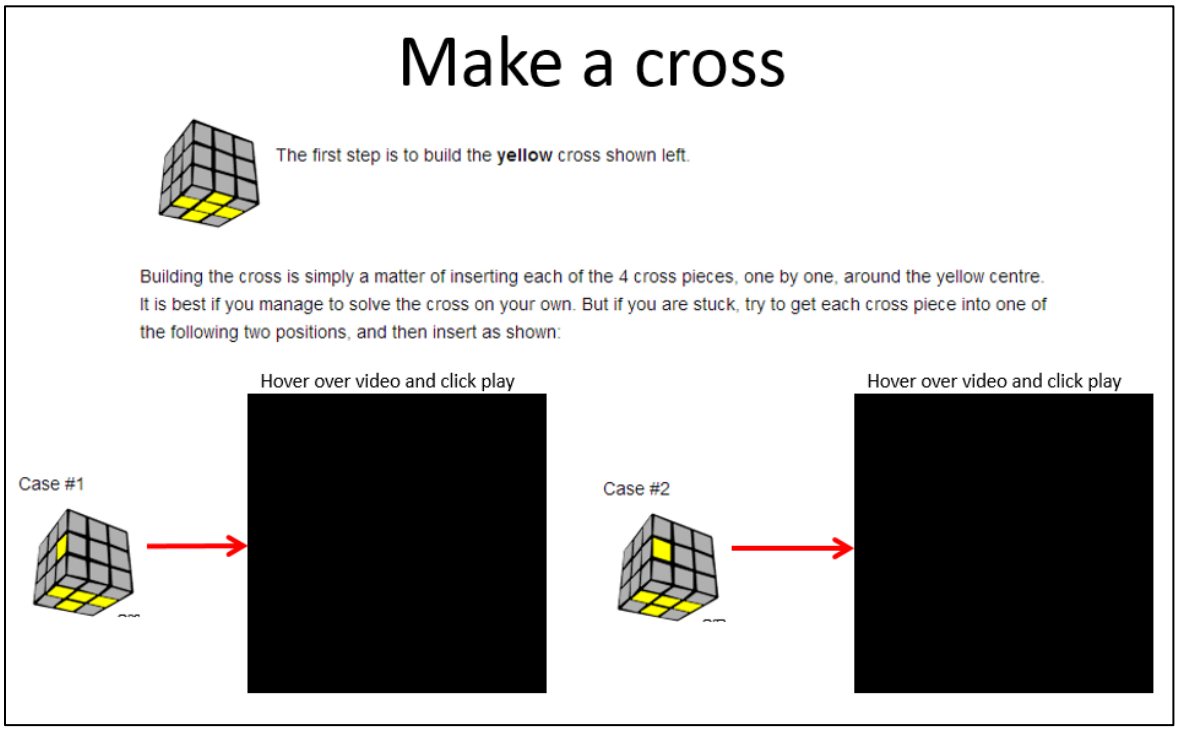
Figure 4. Tutorial interface

The tutorial uses videos with functions similar to those found on the internet (standard fast-forward, rewind, and pause capabilities).

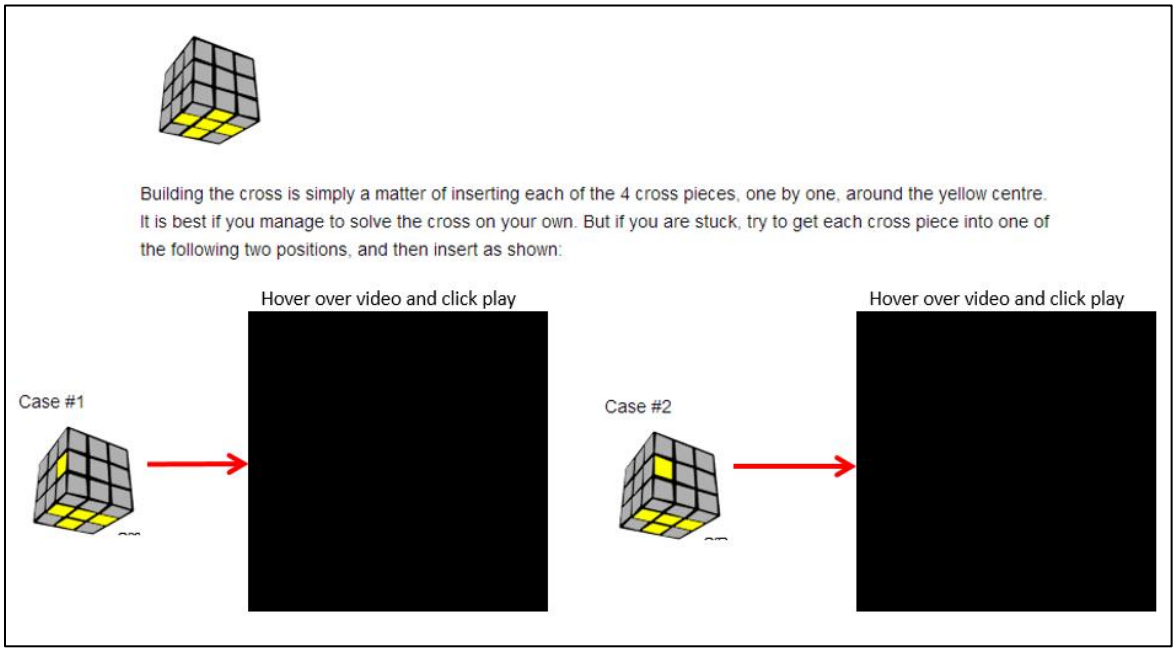
Materials adapted from:
http://mzrg.com/rubik/mech.shtml
http://www.ryanheise.com/cube/beginner.html

APPENDIX K. RUBIK'S CUBE TUTORIAL EXAMPLE SLIDES

K.1 First instruction slide (with subgoal label)



K.2 First instruction slide (without subgoal label)



APPENDIX L. SUBGOAL OUTLINE FOR COMPLETING THE PRIMARY RUBIK'S CUBE TASK (THE "FIRST LAYER")

Rubik's Cube "first layer" subgoals

1. Make a cross by inserting the four yellow cross pieces around the yellow center

2. Match the secondary cross piece colors to their respective adjacent centers

3. Insert the four bottom corners so that the colors of the corner pieces match up with their adjacent cross piece colors. The finished first layer is depicted here:

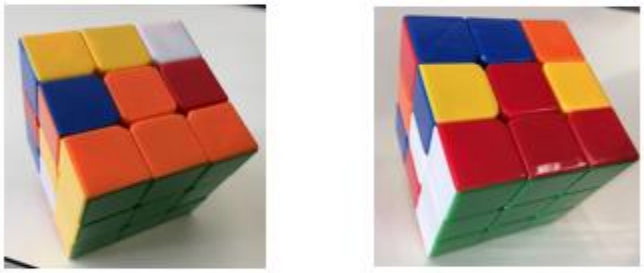
APPENDIX M. RUBIK'S CUBE DOMAIN IMMEDIATE POST-TEST DIRECTIONS

Rubik's Cube post-test

- Make sure that when the timer stops, you have as many pieces in their correct spots as possible

Problem 1 (four minutes): Solve the first layer from the yellow side (same as the task from the instructional materials):

Problem 2 (five minutes): Solve the first layer from the green side (see example of finished product below):



Problem 3 (six minutes): Re-create the following mini USA flag design (don't worry about the colors on the other side, just worry about the top face):

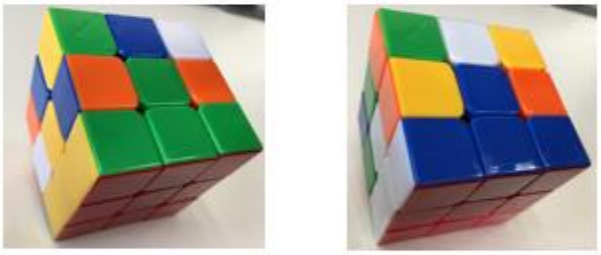
APPENDIX N. RUBIK'S CUBE DOMAIN RETENTION TEST DIRECTIONS

Rubik's Cube retention test

- Make sure that when the timer stops, you have as many pieces in their correct spots as possible

Problem 1 (four minutes): Solve the first layer from the yellow side (same as the task from the instructional materials):

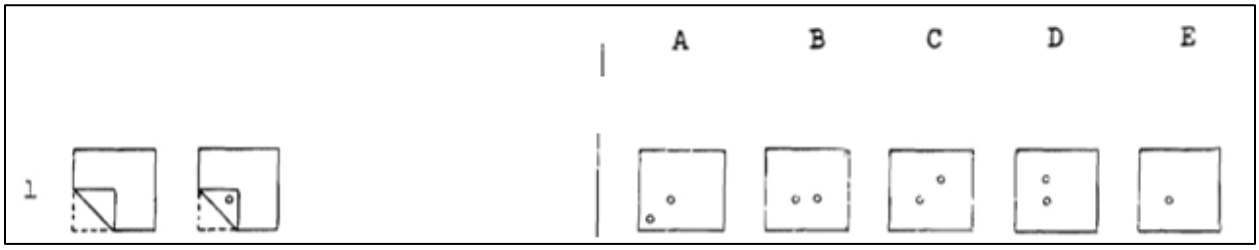
Problem 2 (five minutes): Solve the first layer from the red side (see example of finished product below):



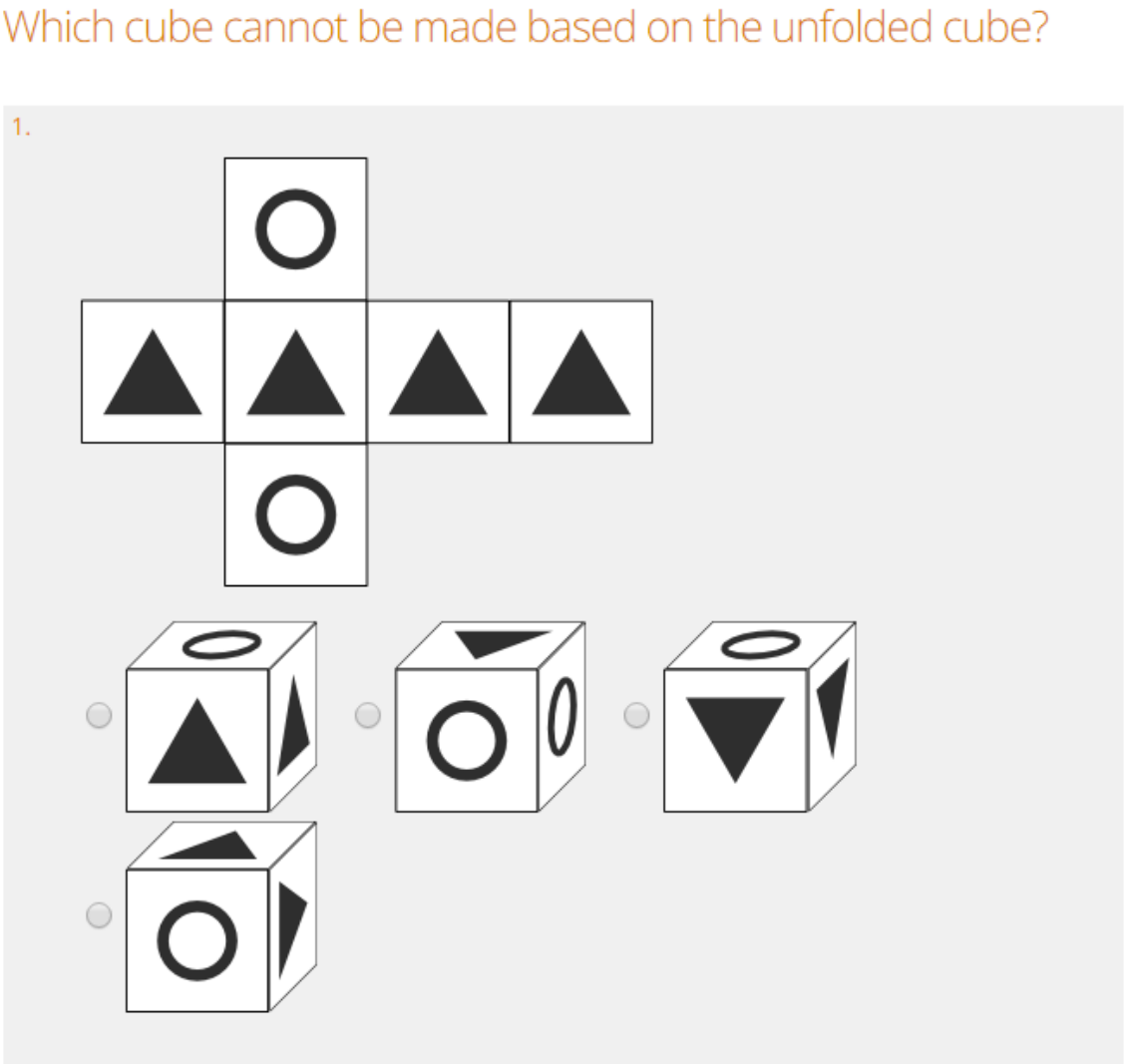
Problem 3 (six minutes): Re-create the following mini France flag design (don't worry about the colors on the other side, just worry about the top face):

APPENDIX O. RUBIK'S CUBE DOMAIN TESTS OF PRE-EXISTING SPATIAL ABILITY

O.1 Example of paper-folding problem



O.2 Example of cube-folding comparison problem



REFERENCES

Anthony, W.S. (1973). Learning to discover rules by discovery. *Journal of Educational Psychology*, *64* (3), 325-328.

Atkinson, R.K., Derry, S., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*. *70* (2), 181–214.

Belenky, D.M. & Nokes-Malach, T.J. (2012). Motivation and Transfer: The role of Mastery-Approach Goals in Preparation for Future Learning. *Journal of the Learning Sciences*, *21* (3), 399-432.

Bjork, E.L. & Bjork, R.A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M.A. Gernsbacher,et al. (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society*. New York: Worth Publishers.

Bjork, R.A. (2013). Desirable difficulties perspective on learning. In H. Pashler (ed.), *Encyclopedia of the mind*. Thousand Oaks, CA: Sage Reference.

Bjork, R.A. & Bjork, E.L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes, Vol. 2* (pp. 35-67). Hillsdale, NJ: Erlbaum.

Blodgett, H. C. (1929). The effect of the introduction of reward upon the maze performance of rats. *University of California Publications in Psychology*, *4*, 113–134.

Bransford, J.D. & Schwartz, D.L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad & P.D. Pearson (Eds.), *Review of research in education*. American Educational Research Association: Washington, DC, 61–101.

Brown, A. & Campione, J. (1994). Guided discovery in a community of learners. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 229-270). Cambridge, MA: MIT Press.

Brown, J.S., Collins, A., & Duguid, P. (1988). Situated cognition and the culture of learning. *Educational Researcher, 18* (1), 32-42.

Bruner, J.S. (1961). The art of discovery. *Harvard Educational Review, 31*, 21–32.

Carlson, R.A., Lundy, D.H., & Schneider, W. (1992). Strategy guidance and memory aiding in learning a problem-solving skill. *Human Factors, 34*, 129-145.

Carpenter, S.K. (2009) Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology – Learning, Memory, and Cognition, 35*, 1563–1569

Carrier, M. & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition, 20*, 633-642.

Carroll, J.M. & Carrithers, C. (1984). Training Wheels in a User Interface. *Communications of the ACM, 27* (8), 800-806.

Catrambone, R. (1998). The subgoal learning model: Creating better examples so that students can solve novel problems. *Journal of Experimental Psychology: General, 127* (4), 355-376.

Chen, D. (2016). *The Role of Struggle and Productive Failure in Learner Assistance*. Unpublished manuscript.

Chen, D. & Catrambone, R. (2016). Facilitating spatial task learning in interactive multimedia environments while accounting for individual differences and task difficulty. *Proceedings of the 38th Annual Meeting of the Cognitive Science Society* (pp. 1925-1930). Austin, TX: Cognitive Science Society.

Chen, D. & Catrambone, R. (2014). Effects of multimedia interactivity on spatial task learning outcomes. *Proceedings of the 58th Annual Meeting of the Human Factors and Ergonomics Society* (pp. 1356-1360). Santa Monica, CA: Human Factors and Ergonomics Society.

Chi, M.T.H., Feltovich, P.J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices. *Cognitive Science, 5*, 121-152.

Chi, M.T.H. (2000). Self-explaining: the dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), Advances in instructional psychology, (pp. 161–238). Mahwah, NJ: Lawrence Erlbaum Associates.

Clark, R.E., Kirschner, P.A., & Sweller, J. (2012). Putting students on the path to learning: the case for fully guided instruction. *American Educator*, *36*, 6-11.

Clement, J. (1991). Non-formal reasoning in science: The use of analogies, extreme cases, and physical intuition. In Voss, J.F., Perkins, D.N., & Siegel, J. (Eds.), *Informal reasoning and education*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Collins, A.M. & Quillian, M.R. (1972). Experiments on semantic memory and language comprehension. In L. Gregg (ed.), *Cognition and learning* (pp. 117-138). New York, NY: Wiley.

Cope, P., & Simmons, M. (1994). Some effects of limited feedback on performance and problem-solving strategy in a logo microworld. *Journal of Educational Psychology, 86* (3), 368-379.

DeCaro, M.S., & Rittle-Johnson, B. (2012). Exploring mathematics problems prepares children to learn from instruction. *Journal of Experimental Child Psychology, 113* (4), 552–568.

Dean, D. & Kuhn, D. (2006). Direct Instruction vs. Discovery: The Long View. *Science Education*, *91*, 384-397.

diSessa A.A. & Sherin, B.L. (2000). Meta-representation: an introduction. *Journal of Mathematical Behavior, 19*, 385-398.

diSessa, A., Hammer, D., Sherin, B., & Kolpakowski, T. (1991). Inventing graphing: Children's meta-representational expertise. *Journal of Mathematical Behavior*, *10* (2), 117-160.

Durkin, K. & Rittle-Johnson, B. (2012). The effectiveness of using incorrect examples to support learning about decimal magnitude. *Learning and Instruction*, *22* (3), 206-214.

Eisenkraemer, R.E., Jaeger, A., & Stein, L.M. (2013). A systematic review of the testing effect in learning. *Paideia*, *23*, 397-406.

Elshout, J.J. (1985). Problem solving and education. Paper presented at the First Conference of the European Association for Research on Learning and Instruction, Leuven, Belgium.

Endres, T. & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*, *6*.

Fiore, S.M., Scielzo, S., Jentsch, F., & Howard, M.L. (2006). Understanding performance and cognitive efficiency when training for x-ray security screening. *Proceedings of the Human Factors and Ergonomics society 50th Annual Meeting* (pp. 2610-2614). Santa Monica, CA: Human Factors and Ergonomics Society.

Gick, M.L. & Holyoak, K.J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Goode, S. & Magill, R.A. (1986). The contextual interference effects in learning three badminton serves. *Research Quarterly for Exercise and Sport. 57*, 308-314.

Goodman, J.S., Wood, R.E., & Hendrickx, M. (2004). Feedback specificity, exploration, and learning. *Journal of Applied Psychology, 89* (2), 248-262.

Gorman, J.C., Cooke, N.J., & Amazeen, P.G. (2010). Training Adaptive Teams. *Human Factors*. *52*, 295-307.

Gray, L.E. (1982). Aptitude constructs, learning processes, and achievement. Unpublished report. Stanford University.

Guzdial, M. (1997, June 16). Constructivism vs. Constructivism vs. Constructionism. Retrieved from: http://guzdial.cc.gatech.edu/Commentary/construct.html

Halamish, V. &Bjork, R.A. (2011). When does testing enhance retention? A distribution- based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 801–812.

Hardiman, P., Pollatsek, A., & Weil, A. (1986). Learning to understand the balance beam. *Cognition and Instruction, 3*, 1–30.

Hiebert, J., & Grouws, D.A. (2007). The effects of classroom mathematics teaching on students' learning. In F.K. Lester (Ed.), *Second handbook of research on mathematics teaching and learning* (pp. 371–404). Charlotte, NC: Information Age.

Jonassen, D. (1991). Objectivism vs. constructivism. *Educational Technology Research and Development*, *39* (3), 5–14.

Kapur, M. (2008). Productive failure. *Cognition and Instruction*, *26* (3), 379-424.

Kapur, M. (2011). A further study of productive failure in mathematical problem solving: unpacking the design components. *Instructional Science, 39*, 561-579.

Kapur, M. (2012). Productive failure in learning the concept of variance. *Instructional Science, 40* (4), 651–672.

Kapur, M. (2014). Comparing Learning from Productive Failure and Vicarious Failure. *Journal of the Learning Sciences, 23* (4), 651-677.

Kapur, M. & Bielaczyc, K. (2011). Classroom-based experiments in productive failure. In L. Carlson, C. Holscher, & T. Shipley (Eds.), Proceedings of the 33rd annual conference of the Cognitive Science Society (pp. 2812–2817). Austin: Cognitive Science Society.

Kapur, M. & Bielaczyc, K. (2012). Designing for Productive Failure. *Journal of the Learning Sciences, 21* (1), 45-83.

Kapur, M., Dickson, L., & Yhing, T.P. (2010). Productive failure in mathematical problem solving. *Instructional Science, 38* (6), 523–550.

Kapur, M., & Lee, K. (2009). Designing for productive failure in mathematical problem solving. In *Proceedings of the 31st Annual Conference Of The Cognitive Science Society* (pp. 2632-7). Austin, TX: Cognitive Science Society.

Karpicke, J.D. & Blunt, J.R. (2011). Retrieval Practice Produces More Learning Than Elaborative Studying with Concept Mapping. *Science, 331*, 772-775

Karpicke, J.D., Butler, A.C., & Roediger, H.L. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory, 17*, 471–479.

Kehoe, C., Stasko, J., & Taylor, A. (2001). Rethinking the evaluation of algorithm animations as learning aids: an observational study. *International Journal of Human-Computer Studies, 54*, 265-284.

Kerr, R. & Booth, B. (1978). Specific and varied practice of a motor skill. *Perceptual and Motor Skills*, *46*, 395-401.

Kirschner, P.A., Sweller, J., & Clark, R.E. (2006). Why Minimal Guidance During Instruction Does Not Work: An Analysis of the Failure of Constructivist, Discovery, Problem-Based, Experiential, and Inquiry-Based Learning. *Educational Psychologist*, *41* (2), 75-86.

Koedinger, K.R. & Aleven, C. (2007). Exploring the assistance dilemma in experiments with cognitive tutors. *Educational Psychology Review*, *19*, 239-264.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General*, *138*, 449–468.

Kulhavy, R.W., & Stock, W.A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, *1* (4), 279-307.

Kulik, J.A., & Kulik, C.C. (1988). Timing of feedback and verbal learning. *Review of Educational Research*, *58* (1), 79-97.

LaPorte, R.E. & Nath, R. (1976). Role of performance goals in prose learning. *Journal of Educational Psychology*, *68*, 260-264.

Lin, X., Hmelo, X., Kinzer, C.K., & Secules, T.J. (1999). Designing technology to support reflection. *Educational Technology, Research and Development*, *47* (3), 43-62.

Loibl, K. & Rummel, N. (2014a). The impact of guidance during problem-solving prior to instruction on students' inventions and learning outcomes. *Instructional Science*, *42*, 305-326.

Loibl, K. & Rummel, N. (2014b). Knowing what you don't know makes failure productive. *Learning and Instruction*, *34*, 74-85.

Margulieux, L.E., Guzdial, M., & Catrambone, R. (2012). Subgoal-labeled instructional material improves performance and transfer in learning to develop mobile applications. *Proceedings of the Ninth Annual International Conference on International Computing Education Research*, 71–78.

Marsh, E.J., & Butler, A.C. (2013). Memory in educational settings. In D. Resiberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 299-317). Oxford: Oxford University Press.

Mathan, S. Koedinger, K.R (2003). Recasting the Feedback Debate: Benefits of Tutoring Error Detection and Correction Skills. In H.U. Hoppe et al. (Eds.), *Artificial Intelligence in Education*, (pp. 13-20). IOS Press.

Mayer, R. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *American Psychologist*, *59*, 14-19.

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *Journal of Experimental Psychology: General*, *131* (3), 349-363.

Mochon, D., Norton, M.I., & Ariely, D. Bolstering and Restoring Feelings of Competence via the IKEA Effect. *International Journal of Research in Marketing*, *29* (4), 363-369.

Moore, J.L. & Schwartz, D.L. (1998). On learning the relationship between quantitative properties and symbolic representations. *Proceedings of the International Conference of the Learning Sciences* (pp. 209-214). Mahwah, NJ: Erlbaum.

Morrison, B.B., Margulieux, L/E., & Guzdial, M. (2015). Subgoals, context, and worked examples in learning computing problem solving. *Proceedings of the Eleventh Annual International Conference on International Computing Education Research* (pp. 21-29). New York, NY: Association for Computing Machinery.

Nelson, T.O. & Leonesio, R.J. (1988). Allocation of Self-Paced Study Time and the "Labor-in-Vain Effect." *Journal of Experimental Psychology*, *14* (4), 676-686.

Patel, V.L., Groen, G.J., & Norman, G.R. (1993). Reasoning and instruction in medical curricula. *Cognition & Instruction, 10*, 335–378.

Peterson, P.L. (1987). Selecting students and services for compensatory education. In B.I. Williams, P.A. Richmond, & B.J. Mason (Eds.), *Designs for compensatory education: Conference proceedings and papers*. Washington, D.C.: Research and Evaluation Associates, Inc.

Reed, S.K., Ackinclose, C.C, & Voss, A.A. (1990). Selecting analogous problems: Similarity versus inclusiveness. *Memory & Cognition*, *18*, 83-98.

Reif, F. & Larkin, J. H. (1991). Cognition in scientific and everyday domains: Comparison and learning implications. *Journal of Research in Science Teaching (Special Issue: Students' models and epistemologies), 28*(9), 733-760.

Renkl, A. & Atkinson, R.K. (2002). Learning from examples: Fostering self-explanations in computer-based learning environments. *Interactive Learning Environments*. *10* (2), 105–119.

Reynolds, R. & Caperton, I.H. (2011). Contrasts in student engagement, meaning-making, dislikes, and challenges in a discovery-based program of game design learning. *Educational Technology Research and Development, 59,* 267-289.

Roediger H.L. & Butler A.C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27.

Rourke, A. & Sweller, J. (2009). The worked-example effect using ill-defined problems: Learning to recognize designers' styles. *Learning and Instruction, 19*, 185-199.

Sander, E. & Richard, J. (1997). Analogical transfer as guided by an abstraction process: The case of learning by doing in text editing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1459-1483.

Schank, R. (1997). *Virtual Learning: A Revolutionary Approach to Building a Highly-Skilled Workforce*. New York, NY: McGraw-Hill

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49,* 31–57.

Schmidt, R.A. & Bjork, R.A. (1992). New Conceptualizations of Practice: Common Principles in Three Paradigms Suggest New Concepts for Training. *Psychological Science, 3* (4), 207-217.

Schwartz, D.L., & Martin, T. (2004). Inventing to Prepare for Future Learning: The Hidden Efficiency of Encouraging Original Student Production in Statistics Instruction. *Cognition and Instruction, 22* (2), 129-184.

Shea, J.B. & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 179-187.

Siegler, R.S. (2002). Microgenetic studies of self-explanation. In N. Garnott & J. Parziale (Eds.), *Microdevelopment: A process-oriented perspective for studying development and learning* (pp. 31–58). Cambridge, UK: Cambridge University Press.

Smith, J.P., diSessa, A.A., & Roschelle, J. (1993). Misconceptions reconceived: a constructivist analysis of knowledge in transition. *Journal of the Learning Sciences*, *3* (2), 115–163.

Snow, R.E. (1982). Education and intelligence. In R.J. Sternberg (Ed.), *Handbook of Human Intelligence*. New York: Cambridge University Press.

Snow, R.E. (1989). Aptitude-Treatment Interaction as a framework for research on individual differences in learning. In P. Ackerman, R.J. Sternberg, & R. Glaser (eds.), *Learning and Individual Differences*. New York: W.H. Freeman.

Soderstrom, N.C. & Bjork, R.A. (2015). Learning versus Performance: An Integrative Review. *Perspectives on Psychological Science*, *10* (2), 176-199.

Sungkhasettee, V.W., Friedman, M.C., & Castel, A.D. Memory and metamemory for inverted words: Illusions of competency and desirable difficulties. *Psychonomic Bulletin and Review*, *18*, 973-978.

Sweller, J. (2010). Element interactivity and intrinsic, extraneous, and germane cognitive load. *Educational psychology review*. *22* (2), 123–138.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*, 257–285.

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, *8* (4), 351–362.

Sweller, J. & Cooper, G.A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. *Cognition and Instruction*, *2*, 59-89.

Sweller, J., Mawer, R., & Howe, W. (1982). The consequences of history-cued and means-ends strategies in problems solving. *American Journal of Psychology, 95*, 455–484.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W.B. (2003). Why Do Only Some Events Cause Learning During Human Tutoring? *Cognition and Instruction*, *21* (3), 209-249.

Vygotsky, L.S. (1978). *Mind in society: The development of higher order psychological processes*. Cambridge, England: Cambridge University Press.

Wineburg, S.S. & Foumier, J.E. (1994). Contextualized thinking in history. In M. Carretero & J. F. Voss (Eds.), *Cognitive and instructional processes in history and the social sciences* (pp. 285-308). Hillsdale. NJ: Erlbaum.

Winograd, T. & Flores, F. (1987). *Understanding Computers and Cognition: A New Foundation for Design*. Boston, MA: Addison-Wesley Publishing.

Young, M.D., Healy, A.F., Gonzalez, C., Dutt, V., & Bourne, L.E. (2011). Effects of training with added difficulties on RADAR detection. *Applied Cognitive Psychology, 25*, 395-407.